How to Evaluate Meta-Models for Software Quality?

Michael Kläs, Constanza Lampasona, Sabine Nunnenmacher
Fraunhofer Institute for Experimental Software Engineering
Fraunhofer-Platz 1, 67663 Kaiserslautern, Germany
{michael.klaes, constanza.lampasona, sabine.nunnenmacher}@iese.fraunhofer.de
Stefan Wagner, Markus Herrmannsdörfer, Klaus Lochmann
Institut für Informatik, Technische Universität München
Boltzmannstr. 3, 85748 Garching, Germany
{wagnerst, herrmama, lochmann}@in.tum.de

Abstract:

The use of appropriate software quality models is crucial for companies to achieve the product quality required to satisfy customer needs. Most current quality models provide little operationalization and lack adaptation guidelines, which limits their usefulness in practice. It has been proposed to use meta-models to specify an explicit structure in order to ensure that quality models conforming to it can be operationalized and adapted by requiring corresponding model elements and modeling constructs. To be applicable in practice, a meta-model needs to be general enough so that existing quality models can be transferred to the new structure provided by the meta-model while preserving the knowledge they contain. This paper presents an empirical approach for evaluating generality as well as its application to a selected meta-model and six industrial quality models. The results show that (1) the proposed meta-model is general enough to model most contents of the industrial quality models, (2) the generality of a meta-model contributes to its perceived ease of use and usefulness, and (3) the evaluation approach is applicable and reflects the perception of quality model experts well.

Keywords

Software quality assurance, quality model, generality, Quamoco, empirical study

1 Introduction

Software quality models (QM) are an important means for supporting quality assurance processes in software development and maintenance. Many QMs (e.g., [15], [3], [7]) and standards (e.g., ISO 9126 [10]) exist. Most of them suffer from deficiencies such as limited operationalization or missing tailoring methods that hamper their acceptance in practice. An explicit meta-model that specifies the structure of QMs is also missing in most models, but could help to address these

deficiencies by defining required model elements and dependencies. In the publicly funded project Quamoco[1], a software QM standard is being developed whose aims are to address existing deficits and to be applicable in practice. This includes a proposal for a quality meta-model (QMM).

Problem Statement. Many software companies use some form of a QM that captures knowledge about quality. To ensure practical applicability and the preservation of this knowledge, any newly proposed QMM needs to be evaluated with respect to its ability to adequately express the contents of existing models. This requires a sufficient degree of generality from the QMM. To the best of our knowledge, however, no ready-to-use concepts exist for evaluating whether the chosen level of generality is appropriate. Therefore, we operationalize the concept of QMM generality and develop the needed measurement instruments and procedures for the evaluation.

Objective. The main objective is to develop an evaluation approach for QMMs in order to analyze and improve the QMM developed in Quamoco. This evaluation approach, however, is intended to be generally applicable to any QMM.

Contribution. The contribution is threefold: (1) We present a first version of a QMM, which provides the concepts for specifying and evaluating software quality while addressing limitations of existing QMs. (2) We provide an *empirical evaluation approach* for characterizing and evaluating the generality of QMMs. (3) We present the results of a series of studies where the approach was applied to six industrial QMs. They show the applicability of the approach and correspondence of the measurement results with the subjective evaluations of QM experts. Further, they provide an initial evaluation of the proposed QMM and identify improvement potentials for future work on QMMs.

2 Related Work

Quality Meta-Models. Kitchenham et al. [11] first proposed separating structure and content when modeling software quality. A meta-model defines the structure; the content is a model conforming to the meta-model. To evaluate their QMM, they applied it to model the quality requirements for Telescience subsystems. Since then, a number of meta-models for modeling software quality have been proposed. Marinescu and Ratiu [13] present the factor-strategy QMM to quantify the quality of object-oriented design. It associates quality-influencing factors with detection strategies for measurement. They applied the QMM to assess two industrial systems. The activity-based QMM of Deissenboeck et al. [6] had a strong influence on the Quamoco QMM. It models the impact of quality factors on activities. To demonstrate its applicability, it was applied in an industrial context to

[1] http://www.quamoco.de/

model quality guidelines for the maintainability of Simulink models. Additionally, they employed it to model usability as another aspect of software quality [16]. Kläs et al. [12] defined an abstract QMM with high-level concepts required for specific QM application purposes. The concepts' appropriateness was evaluated by studying approx. 80 QMs. These QMMs are either abstract or only have been shown to be general enough to describe software quality in a small number of case studies. In contrast, we propose a method for systematically determining the generality of a QMM with the purpose of integrating several, possibly heterogeneous QMs.

Evaluation of Meta-Models. There is also related work on evaluating meta-models in general. Chen et al. examine which factors influence the success of programming languages [4]. Their results indicate that *generality* is the third most important factor, behind machine independence and extensibility. Other work focuses on empirically evaluating the usefulness of metrics for assessing the quality of meta-models. Bajaj [1] conducted experiments to evaluate the use of metrics for the readability of data models; Genero et al. [8] conducted experiments to evaluate the use of metrics for the understandability or modifiability of UML class diagrams. Most closely related are approaches that evaluate a meta-model by analyzing the mapping of the meta-model to an ontology. Guizzardi et al. [9] present a method for analyzing the mapping of the meta-model to a domain ontology with regard to *domain appropriateness*, i.e., the suitability of a modeling language for modeling concepts of a certain domain. Opdahl and Henderson-Sellers [15] present a method for analyzing the mapping of the meta-model to the Bunge-Wand-Weber model in terms of *construct overload, construct redundancy, construct excess,* and *construct deficit*. Our approach requires the quality experts to transcribe several QMs to the evaluated QMM. Using a questionnaire, they can systematically analyze the mapping.

3 The Quamoco Meta-Model

This section describes the QMM whose generality we evaluated. It was developed in the Quamoco project and defines a structure that a QM needs to conform to. We use conforming QMs to specify and evaluate the quality of software products.

3.1 General Concepts

Figure 1 visualizes the constructs and the relationships provided by the QMM. The QMM can be logically separated into two parts: *specification* and *evaluation*.

Specification. The purpose is to specify the quality of a software product in a qualitative manner. The idea of specifying quality using QMs dates back to the early 1970s with the QMs of McCall [14] and Boehm [3]. In McCall's QM, *factors* and *quality aspects* are separated for the first time. This was extended by

Dromey [7], who replaced factors with structural elements of software and characteristics thereof. The separation of software characteristics and influenced quality aspects is a central concept in QMs. Deissenboeck et al. [6] also rely on this principle, but describe the influence of factors on the cost of maintenance activities in order to avoid more diffuse concepts like maintainability, understandability, etc. In our QMM, the separation is represented by *factors* having an *impact* on *quality aspects*. Factors are defined by *entities* of the system and *properties* characterizing them.

Evaluation. Evaluation defines software product quality in a quantitative manner. Concrete *measures* and *evaluations* are described, e.g., through thresholds or expert judgment. A measure quantifies a factor. An evaluation is performed on the impact rather than the factor, as the measurement of a factor usually has different impacts on different quality aspects. For instance, the structuredness of the system has to be evaluated differently for maintainability and performance. A monolithic system usually performs better but is harder to maintain. The evaluation of a quality aspect can be performed based on the evaluation of the impacts that influence the quality aspect as well as on evaluations of lower-level quality aspects.

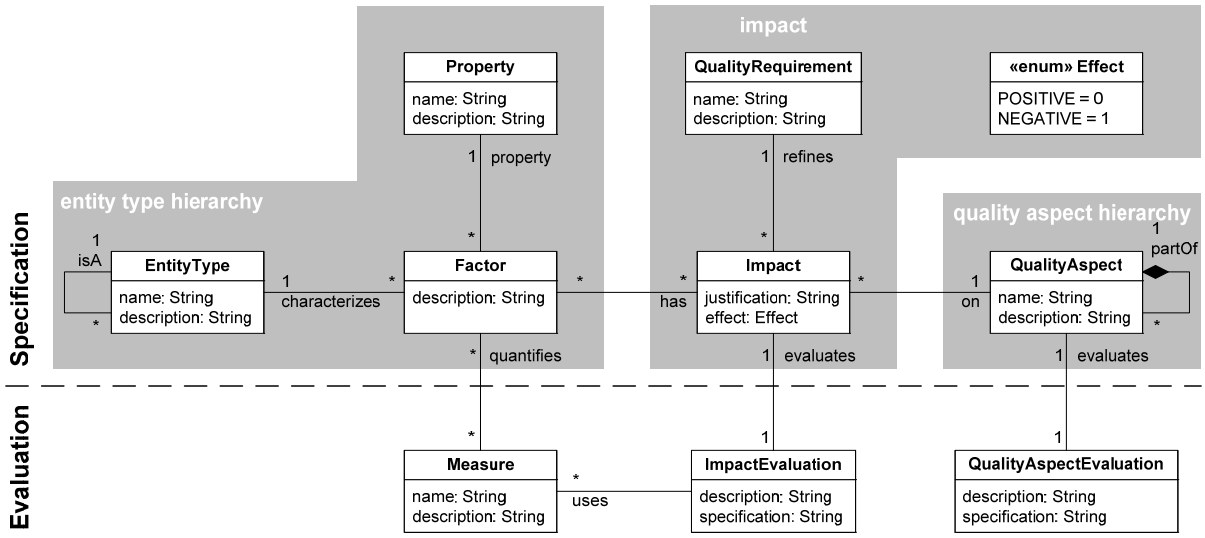

Figure 1: Quamoco meta-model for specifying and evaluating software quality

3.2 Quality Specification

The Quamoco meta-model provides the following constructs for specifing quality.

Quality Aspect. A quality aspect describes a focus that is addressed by the QM, such as maintainability. It can be decomposed into sub-aspects, resulting in the so-called *aspect hierarchy*. Since different stakeholders have different orthogonal perceptions of quality, several quality aspect hierarchies are allowed and the meta-model does not prescribe a specific one. The QMM can therefore accommodate different quality aspects, such as the quality attributes of ISO 9126 [10] (i.e., *reli-*

ability, *usability*, etc.) for communicating evaluation results to managers, activities for analyzing financial value and costs, or technical topics to be communicated to developers.

Factor. A factor is a circumstance or fact in the software product and its environment that has an influence on product quality. Example factors are the *redundancy of source code*, i.e., the well-known phenomenon of code cloning, or the *consistent usage of fonts in UI widgets*. In contrast to quality aspects, factors describe on a lower level what is important for software product quality.

Impact. An *impact* defines an influence of one or more factors on a quality aspect. The effect of the influence can be positive or negative. Each impact needs to specify in an explicit justification why the factor has an impact on the quality aspect. This description of the rationale helps to ensure that the model contains only *relevant* impacts. Examples include the *negative impact of code redundancy on modification,* which describes that redundant source code is hard to modify, or the *positive impact of widget font consistency on usability*, which describes why consistent usage of fonts in a UI improves the usability of a software product.

Entity Type. An entity type is a type of element that is part of the software product, or that is a resource required during development, maintenance, or use of a software product. It describes to what type of entity a factor is related. A generalization/specialization relation between entity types forms the *entity type hierarchy*. Example entity types that describe product artifacts are *system, code, test case,* or *widget font*. Entity types that describe resources used during development are *developer, version management system, debugger,* or *review process*.

Property. Properties are general attributes that characterize what a factor describes about an entity type. Properties enable clear separation between the parts of a software product and their attributes that influence quality. The separation of entity types and properties is based on Kitchenham et al. [11], who state that entities "are the objects we observe in the real world" and attributes are "the properties that an entity possesses". We reuse properties, i.e., a single property describes entities in a number of factors. Furthermore, an entity type can be characterized by a number of properties, leading to different factors. Example properties are *consistency, conformance, conciseness, redundancy, superfluousness* or even simple *existence,* which describes whether an element of the entity type exists.

Quality Requirement. Following ISO 9126 [10], quality requirements specify the required level of quality. A quality requirement is refined into a set of factors with an impact on quality aspects. A common quality requirement is "The code shall be easy to understand". This could be concretized in the QM, for example, by associating the factors *conciseness and consistency of identifiers* and *appropriateness of comments* and their positive impacts on the quality aspect *program comprehension*.

3.3 Quality Evaluation

The meta-model provides the following constructs for evaluating quality.

Measure. A measure quantifies a factor by defining a method for measuring it using a certain scale. Measures can be reused for the quantification of different factors. Furthermore, a factor can be quantified by different measures, enabling a more profound evaluation. Examples of measures are *clone coverage, number of use cases, test coverage,* or *number of architecture violations*.

Impact Evaluation. We evaluate the impact of a factor on a quality aspect. The QM specifies exactly one evaluation for each impact. The impact evaluation is based on the measures associated with the factors having the impact on the quality aspect. The impact can use all measures defined for its factors. For these measures, we aggregate the measurement data for all entities of the specific entity type. For example, we can define an impact evaluation for the negative impact of code redundancy on modification by mapping possible results of the single measure *clone coverage* onto school grades.

Quality Aspect Evaluation. An aspect evaluation evaluates a quality aspect. The QM specifies one evaluation for each quality aspect. The aspect evaluation is based on the evaluation of the impacts that influence the quality aspect and on lower-level quality aspects. Quality aspect evaluations do not evaluate measurement results, but results on a defined evaluation scale. For instance, we evaluate the maintainability of a system based on the weighted average across several impact evaluation results.

4 Meta-Model Evaluation Approach

In this section, we propose an empirical approach to evaluating the generality of a QMM. First, we present the objectives addressed by the approach, then we describe the design and implementation of studies using the approach.

4.1 Objectives and Scope Definition

Our approach aims to evaluate the *generality* of a QMM. We understand by generality the property of a QMM that allows it to appropriately describe existing QMs. In our experiences, it is a key characteristic that contributes to QMM acceptance in practice.

Objectives. The objectives addressed by the approach are twofold. (1) It should test if a sufficient level of generality is present in the evaluated QMM, enabling the QM experts to use the QMM to transcribe existing models. (2) It should help to identify improvement potentials for QMMs.

We define the *objectives* using the goal/question/metric template [17]:

Analyze a *quality meta-model* for the purpose of *characterization and evaluation* with respect to its *generality* [Quality Focus] from the point of view of the *QM developer and maintainer* in the context of a set of *organizations applying QMs*.

Definition of Scope. The structure of a QM (the underlying QMM) depends on the *application purposes* to be supported by the model. For instance, a model supporting the *specification* of quality usually requires different conceptual constructs than a model intended for the *prediction* of product quality [12]. Hence, the evaluation should only consider those QM constructs that are relevant for the purposes supported by the considered QMM. If the QMM, for example, explicitly does not support predicting the final product quality as a purpose, QM constructs that are only relevant for prediction do not have to be realizable by the QMM. Some QMs are too extensive to be transcribed completely in a generality evaluation study with reasonable effort. To study generality problems relevant for these QMs, we select and model a *representative part* (excerpt) of the QM.

4.2 Design and Implementation

Operationalization of the Construct Meta-Model Generality. To obtain an empirically profound statement about the generality of a QMM, we have to use the QMM to bring existing QMs into the structure defined by it. By doing this, we indirectly evaluate the generality of the QMM by comparing the transcribed QM with the original QM. To find suitable measures, we use the GQM approach [2]. We derive relevant measures through the use of questions supporting the measurement goal, in this case our study objectives. In Table 1, we present all derived Questions related to the quality focus *generality* and the data to be collected. We collect both *quantitative* (i.e., Measures, closed questions) and *qualitative* (i.e., Open Information, interpretative, not numerical) data.

Q1	**Is the quality meta-model sufficiently general to describe the selected part of the existing (original) QM?**
-.1.1	Does the QM expert consider the QMM as sufficiently general to describe the selected part of the existing QM?
M1	***Perceived QMM generality:*** Subjectively evaluated (ordinal scale*)
-.1.2	Why does the QM expert consider the QMM as (not) sufficiently general?
OI	***Generality text:*** Argumentation for subjectively rated generality
Q1.2	Is the transcribed model complete? *Note*: If the transcribed model contains the same relevant concepts as the original model for the supported purposes, then it is said to be complete (Figure 2).
-.2.1	Does the QM expert consider the transcribed model as complete?
M2	***Perceived QM completeness:*** Subjectively evaluated (ordinal scale*)
-.2.2	Why does the QM expert consider the transcribed model as (not) complete?
OI	***Completeness text:*** Argumentation for subjectively rated completeness
-.2.3	What percentage of the 10 most important concepts (e.g., entities, attributes, and relationships) in the original QM can be mapped to the QMM?
M3	***% of concepts supported:*** Percentage of the selected QM concepts supported (ratio scale)
Q1.3	Is the transcribed model understandable? *Note*: If the transcribed model can be interpreted in a reasonable manner and does not become too complex, it is said to be understandable (Figure 2).

-.3.1	Does the QM expert consider the transcribed model as understandable?
M4	***Perceived QM understandability:*** Subjectively evaluated (ordinal scale*)
-.3.2	Why does the expert consider the transcribed model as (not) understandable?
OI	***Understandable text:*** Argumentation for subjectively rated understandability
-.3.3	Is the complexity of the transcribed model increased by distributing concepts of the original QM over several QMM concepts (increasing the number of instances and their relationships)?
M5	***# QM concepts split:*** Number of selected concepts in the original model that are split into two or more elements in the transcribed model (absolute scale) ***% of QM concepts split:*** Ratio between # QM concepts split and total number of selected concepts of the original model (ratio scale)
M6	***# unused QMM concepts:*** Number of concepts in the QMM that are not used in the transcribed model (absolute scale) ***% of unused QMM concepts:*** Ratio between # unused QMM concepts and total number of concepts in the QMM (ratio scale)
Q1.4	Are the semantics of the selected excerpt of the original model retained by the transcribed model? *Note*: The transcribed model is said to keep the semantics of the original model if it can be interpreted in only one way, i.e., if it can be used for the same purposes as the original model (Figure 2).
-.4.1	Does the QM expert think that the semantics of the selected excerpt of the original model are retained?
M7	***Perceived retained semantics:*** Subjectively evaluated (ordinal scale*)
-.4.2	Why does the QM expert think that the semantics of the selected excerpt of the original model are (not) retained?
OI	***Semantic text:*** Argumentation for the subjectively rated semantic retention
-.4.3	What percentage of the selected original QM concepts that are supported by the QMM can be described explicitly?
M8	***% of concepts explicitly supported:*** Percentage of relevant concepts explicitly supported (ratio scale)
Q1.5	Were there concrete problems (issues) during transcription of the QM? *Note*: Issues are specific difficulties and problems in representing specific elements of the QM content in the structure of the QMM.
-.5.1	What are the issues that were detected during transcription of the original QM?
OI	***Issues:*** List of issue descriptions (subjective, enumeration)
-.5.2	What is the type of the issue detected by transcribing the selected excerpt of the original QM?
M9	***Type of issue:*** Subjective classification of the issue (nominal scale: {Completeness, Complexity, Semantic, Other})

Table 1: Operationalization of the quality focus (Meta-model generality)

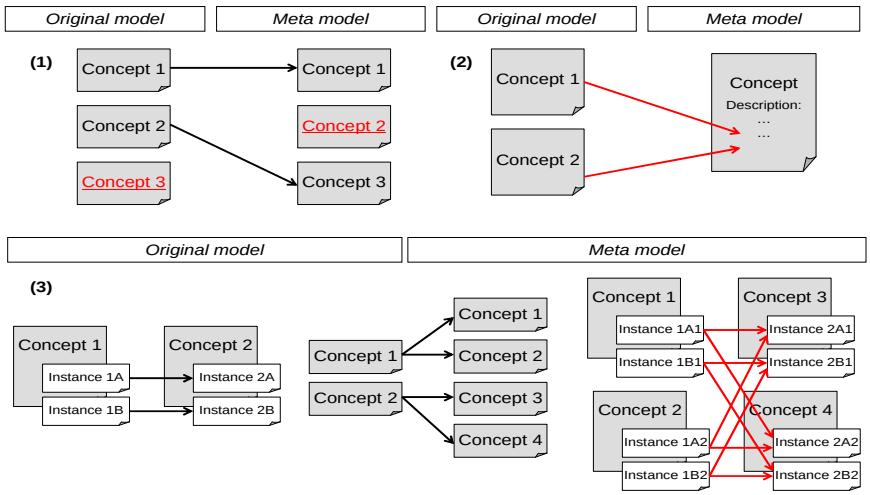

Figure 2: (1) Concept 3 of the original QM cannot be mapped to the QMM and QMM Concept 2 is unused. (2) Loss of model semantics by mapping multiple concepts of the original QM to the description part of the same QMM concept. (3) The concepts of the original QM are unnecessarily distributed over several QMM concepts, increasing the number of instances and their relationships.

Design. The approach focuses on characterizing and evaluating the generality of one selected QMM based on a set of existing QMs used in companies from different domains. We are not primarily interested in how a specific context/variation factor (e.g., experience of subject) may affect the transcription of a QM. Therefore, the strategy we propose consists of a set of small *studies* with as much diversity as possible in the considered study *objects* and *subjects*. This leads to a design with *n* studies, where each considers one different QM (*object*) and is performed by one different QM expert (*subject*).

Training the subjects in using the tool used for the transcription and in understanding the QMM is important, since it is not the understandability of the QMM or the ease of use of the tool that are be evaluated, but the generality of the QMM.

Setting and timing. The studies occur independent of time and place across all subjects.

Overall procedure. The subjects use the QMM to transcribe an excerpt of their QM. Each subject transcribes the QM in which he is an expert, generates an issue list that documents the transcription problems, and evaluates the QMM generality.

Step 1: Each subject selects a representative part of the original QM and documents its selection. Next, the subject identifies all relevant QM concepts that are to be mapped and documents them in the transcription form.

Step 2: The subject uses the QMM, the transcription form (now containing QM concepts to be mapped), and the modeling tool to transcribe the QM. The output consists of the mapping documentation, an issue list, and the transcribed model.

Step 3: The subjects provide their subjective evaluation of the QMM's generality and relevant context information by answering a questionnaire provided to them.

Selection of Study Subjects. If possible, the subjects should be experts with respect to the QM they transcribe and the meta-model. In order to find an appropriate transcription of the original QM, the subject needs a good understanding of the original model and of the meta-model. Only subjects with a profound understanding of the original QM and sufficient knowledge about the meta-model can evaluate the transcribed QM with respect to its completeness, understandability, and retained semantics. A lack of knowledge regarding the QM may lead to overlooking potential problems. Misunderstanding the QM or the meta-model may also lead to inappropriate modeling, or erroneously identified problems. Therefore, the subjects' experience may be an important confounding factor influencing the result of our study and should be explicitly documented.

Instrumentation and Material. Each subject receives a form containing an introduction to the study, a table for documenting the mapping of concepts and problems with this mapping, a table for documenting the transcription issues, and a questionnaire for documenting their subjective evaluation and relevant context in-

formation (Table 2). In addition, the questionnaire asks the subject about the *perceived usefulness* and *perceived ease of use* of the QMM using the standardized questions taken from the technology acceptance model [5].

Q2	**Do characteristics of the subjects influence the results of the study?**
-.1	What is the experience of the subject with the QMM?
M10	***Experience with QMM:*** Subject's experience with the QMM (subjective, ordinal): 3: The person was involved in the development of the meta-model; 2: The person has already used the meta-model for modeling some quality aspect; 1: The person knows and understands the meta-model well, but did not develop or use it; 0: The person has no previous experience with the meta-model.
-.2	What is the experience of the subject with the original model (the model being transcribed)?
M11	***Experience with QM:*** Subject's experience with the original QM (subjective, ordinal): The person … 3: was involved in the development of the model; 2: has already used the model for evaluating some quality aspect; 1: knows the model well, but did not develop or use it; 0: has no previous experience with the model.
-.3	What is the experience of the subject with the tool used to transcribe the model?
M12	***Experience with tool*****:** Subject's experience with the modeling tool (subjective, ordinal): The person … 3: was involved in the development of the tool; 2: has already used the tool to model same exemplary quality aspect; 1: knows the transcription tool well, but did not develop or use it; 0: has no previous experience with the transcription tool.
Q3	**Do characteristics of the object or instrumentation influence the results of the study?**
-.1	Do the subjects consider the transcribed part of the QM as appropriately selected and sufficiently large to lead to representative study results?
M13	***Appr. selection of QM part:*** Subjectively evaluated appropriateness of QM part (subjective, ordinal scale*) ***Appr. size of QM part:*** Subjectively evaluated sufficient size of QM part (subjective, ordinal scale*)
-.2	Were there problems with the tool (hindering the transcription)?
M14	***Appropriateness of tool:*** Subjectively evaluated appropriateness of tool (subjective, ordinal scale*)
-.3	Were there problems identifying relevant constructs of the original QM?
M15	***Possible to ident. QM concepts:*** Subjectively evaluated (subjective, ordinal scale*)
Q4	**Was there anything else that the expert considers to have had influence on the QMM generality?**
OI	***Other influences text:*** Description of other things that may influence generality (qualitative).

Table 2: Operationalization of variation factors

Hypotheses. We propose testing whether the majority of the subjects agree that the meta-model is sufficiently general (Q1.1.1), as well as that the transcribed QM is complete (Q1.2.1), understandable (Q1.3.1), and retains the original semantics (Q1.4.1).

H_{A1}: If the subjects use the QMM, they agree that the QMM is sufficiently general (i.e., H_{A1}: $\mu(M1) > 1.5$; H_{01}: $\mu(M1) \leq 1.5$).

H_{A2}: If the subjects use the QMM, they agree that the transcribed model is complete (i.e., H_{A2}: $\mu(M2) > 1.5$; H_{02}: $\mu(M2) \leq 1.5$).

H_{A3}: The subjects agree that the transcribed model is understandable (i.e., H_{A3}: $\mu(M4) > 1.5$; H_{03}: $\mu(M4) \leq 1.5$).

H_{A4}: If the subjects use the QMM, they agree that the transcribed model retains the semantics, meaning that the model transcribed can be used for the same purposes as the original model (i.e., H_{A4}: $\mu(M7) > 1.5$; H_{04}: $\mu(M7) \leq 1.5$).

* where the Hox are the null hypotheses we plan to reject and μ is the measure of central tendency (in this case, the median)

Due to the scale of measurement, all hypotheses should be tested with a one-tailed Wilcoxon signed rank test, the non-parametric correspondent to the Student t-test. As significance levels, we propose using alpha = 0.1 due to the typically low number of data points. Besides testing the hypotheses, an important aspect will be to learn from the study to improve the QMM. The lists with the classified issues and the qualitative information (OIs in Table 1) are collected for this purpose.

5 Empirical Study

In this section, we present an empirical study that employs the presented evaluation approach (Section 4) for evaluating the Quamoco QMM (Section 3).

5.1 Study Context and Execution

Next, we provide information about the study context and its execution, especially regarding which QMs were used for the evaluation and who participated.

Study Objects. The transcribed QMs cover different domains, were defined with different application purposes in mind, and focus on different quality aspects.

- The *Software Cockpit QM* is used by Capgemini AG for software quality controlling and is applied company-wide in custom software development projects.
- The *itestra QM* is thought to be universally used but focuses on business information systems and financial aspects of quality.
- *Product Standards and Q-Index*: The SAP-specific QM contains the requirements from the Product Standards and measures from the Q-Index. It is applied company-wide for different kinds of software development projects.
- The *SPQR QM* is used by Siemens AG and focuses on the internal code quality of embedded systems. It is used for assessment and controlling based on a significant set of automatically collected code measures.
- The *MAN QM* is an activity-based QM used by MAN Nutzfahrzeuge and was developed by TU München. It focuses on the maintainability of embedded software systems in the automotive domain.
- The *ISO 9126 QM* refines internal, external, and quality in use in quality characteristics and sub-characteristics and provides a list of associated measures.

Study Subjects. Except for the MAN QM and ISO QM, the transcription was done by a professional working in the specific company as a quality manager or in a comparable role who knew the model to be transcribed very well. The MAN QM was transcribed by a researcher who participated in the development of the QM. The ISO QM was transcribed by a researcher who had previously worked with the ISO model.

Scope. The evaluated QMM was developed with the assessment of the product quality and the definition of checkable quality requirements in mind. Components of the existing QMs that are not required for these scenarios and cannot be transcribed in the QMM do not imply a limitation of its generality.

Execution. The study followed the approach defined in Section 3. However, no explicit training did occur because we assumed that all subjects had sufficient experience with the tool and the QMM.

5.2 Study Results and Interpretation

First, we present data we collected that describe the study context (Q2-3). Next, we provide the study results for the QMM (Q1) and discuss outliers based on the plain text answers provided by the study subjects. Finally, we test the hypotheses and analyze the correlations between the study results and factors potentially influencing them.

Variation Factors (Context). According to questions Q2 and Q3, specific abilities of the subjects, properties of the transcribed QM part, and the instrumentation used in the study may influence the study results. Therefore, information about these factors was collected explicitly. As shown in Figure 3, the experience of the subjects with the original QM and the QMM was, in general, high (median = 2.5), which was demanded by the evaluation approach. The majority also had experience with the tool used for the transcription (median = 2). All participants considered the size and representativeness of the transcribed QM part as appropriate for obtaining reliable results (min = 2). Only one subject answered that the provided tool hindered the transcription of the original QM, and all subjects (strongly) agreed that they could identify the most relevant concepts in the original QM (min = 2). Interpretation: We see no major threats to the validity of the results caused by these factors.

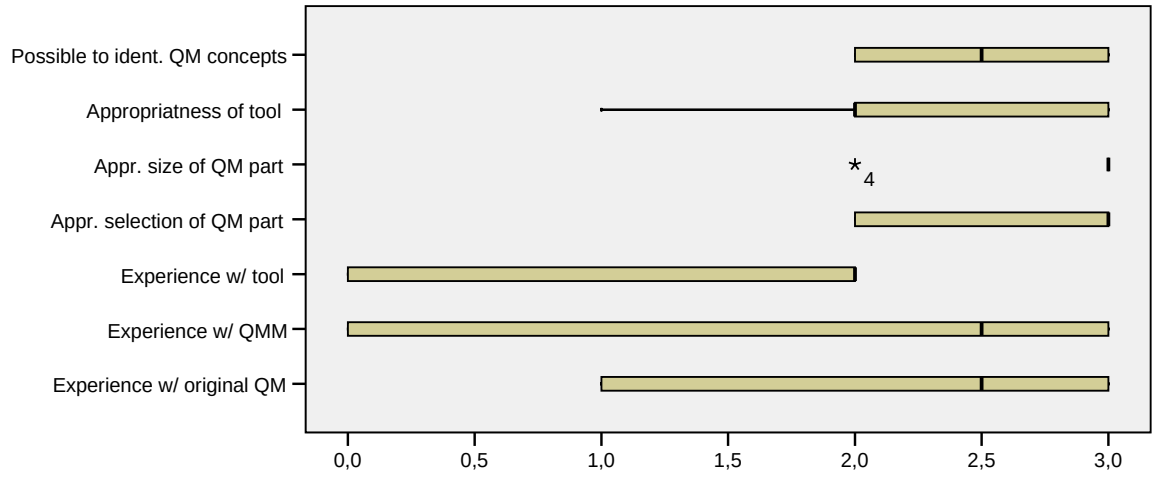

Figure 3: Appropriateness of subjects, objects, and tools
(0: strongly disagree / low to 3: strongly agree / high)

Results of the Subjective Evaluation. We measured the perceived generality of the QMM as well as the perceived completeness, understandability, and retained semantics of the transcribed QM (Figure 4). The majority of the subjects agreed that the QMM is sufficiently general, and the transcribed QM is complete, understandable, and retains the semantics of the original model (median = 2).

There are some outliers that suggest a more detailed look at the provided free-text justifications. The participant disagreeing that the QMM generality is appropriate could not find model elements to provide useful information for the measures that are needed for practical purposes, and perceived that the QMM is overly complex. The participant who rated the understandability of the transcribed model with *strongly disagree* explained that the original QM was not simple, but the transcribed QM was even more difficult to understand when browsed in the provided tool. The free-text justification of the participant who rated the retained semantics with *strongly agree* showed that the person misunderstood the question and rated the degree of semantics of the QMM and not the degree of retained semantics of the transcribed QM. Actions: We excluded the strongly agree answer for semantics before further analysis.

Hypotheses tested using one-tailed Wilcoxon signed rank tests with alpha=0.1

- H_{A1} (Generality, p=0.09) was accepted,
- H_{A2} (Completeness, p=0.38), was not accepted,
- H_{A3} (Understandability, p=0.39) was not accepted,
- H_{A4} (Semantics, p=0.01) was accepted with high significance.

Interpretation: Most probably caused by the low number of cases, we obtained a highly significant result only for the factor *retained semantics*. However, we consider QMM generality at the 0.1 level of significance and the descriptive results (Figure 4) as additional indicators of good generality of the evaluated QMM.

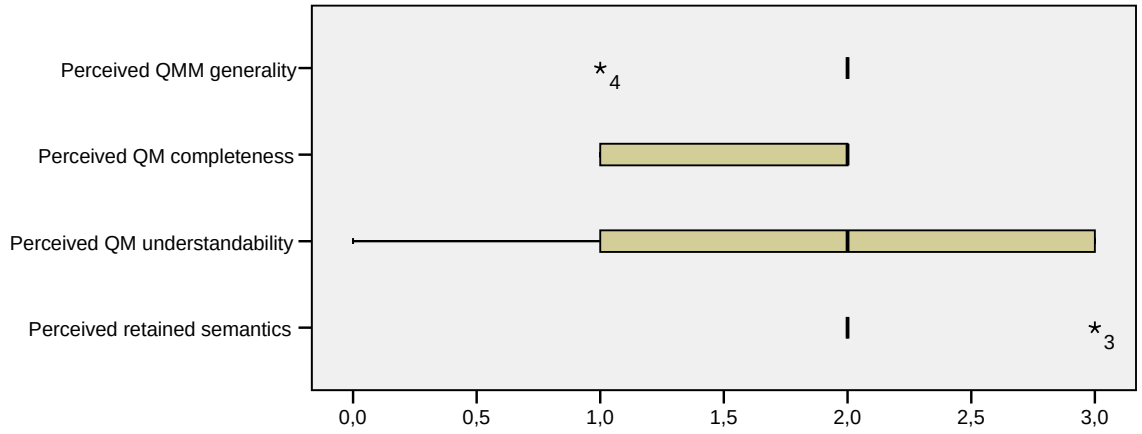

Figure 4: Subjective evaluation results on an agreement scale (0: strongly disagree to 3: strongly agree)

Measurement Results. The measures collected for the relative number of relevant original QM concepts *supported*, *split*, and *explicitly supported* by the QMM (Figure 5) show that 90-100% of the original concepts could be transcribed using the QMM. Mostly, only few concepts in the original QM were separated into multiple concepts by the QMM (~10%). In the majority of the cases, more than 70% of the original concepts were explicitly transcribed by the QMM. The participant who used only 30% of the QMM concepts mentioned that the transcribed model contains a lot of poorly understood elements caused by the concept *factor*.

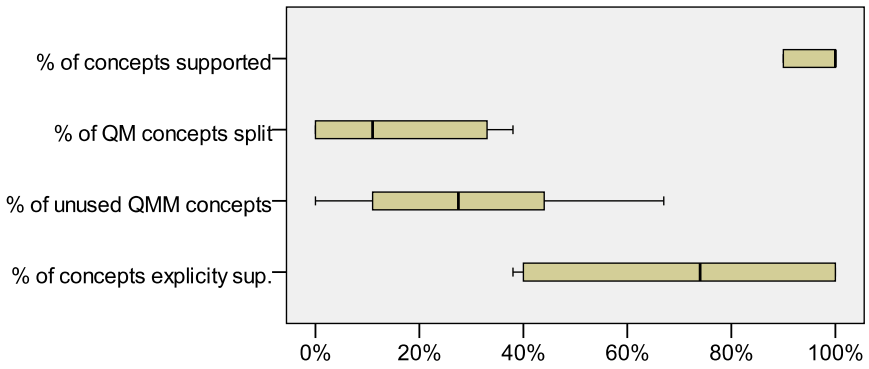

Figure 5: Results for measures derived to describe generality

Impact of Variation Factors. We analyzed the correlation (Spearman's rho) between the subjective generality evaluations and the subject-, object-, and tool-related factors potentially influencing the evaluation results (see Table 3).

Interpretation: (1) The higher the experience with the QM and QMM, the more concepts the participant missed (perceived completeness). This confirms our recommendation to use participants with experience regarding the original QM and the QMM. (2) If the participant was not sure whether the part of the QM he/she transcribed was appropriate, he/she provided a lower generality rating.

Spearman's rho > 0.5: gray, p<0.10: bold	Perceived QMM generality	Perceived QM completeness	Perceived QM understandability	Perceived retained semantics
Experience w/ original QM	-0.42	-0.67	-0.25	N/A
Experience w/ QMM	-0.42	-0.67	0.11	N/A
Experience w/ Tool	N/A	-0.41	0.15	N/A
Appr. selection of QM part	0.63	0.25	-0.11	N/A
Appr. size of QM part	**1.00**	0.61	0.36	N/A
Appropriateness of tool	0.14	-0.34	0.64	N/A
Possible to ident. QM concepts	-0.45	-0.71	0.50	N/A

Table 3: Correlation between context/variation factors and perceived generality

Measure of Improvement Potential (Completeness). In order to identify and measure the degree of improvement potential, we used a table that categorizes the original QM concepts with respect to their relevance as rated by the participants

and whether and how they are addressed in the QMM (see Table 4). Overall, 96% of the original QM concepts can be described (explicitly or implicitly) based on the new QMM. The major improvement potential can be seen in the concretization of the QMM, resulting in explicit support for most of the important and very important concepts of existing QMs.

Original QM Concepts	Not supported	Implicitly supported	Explicitly supported
Less important	2% (Mi)	2% (No)	0% (No)
Important	2% (Ma)	9% (Mi)	11% (No)
Very important	0% (C)	22% (Ma)	52% (No)
		No improvement potential	**65%**
		Minor improvement potential	**11%**
		Major improvement potential	**24%**
		Critical improvement	**0%**

Table 4: Improvement potentials – Unsupported QM concepts

Unused QMM Concepts. The most frequently used concepts were *quality aspects* and *measures* (100%). The concepts *factors, impacts, entity types, properties,* and *requirements* were also used frequently (67%). The concepts *impact evaluation* and *quality aspect evaluation* were used only by a limited number of transcribed QMs (33%). These concepts are not needed in the quality specification use case.

Qualitative Improvement Information (Issues). In Q1.5, the participants could name and classify issues that they had during transcription of the original QM. (1) *Completeness*: The participants listed a number of concepts and attributes that they missed in the QMM. They missed an explicit product model on which to perform aggregations for evaluation. (2) *Complexity*: The participants named a number of concepts that they found superfluous or difficult to use. It was not always easy for them to decompose a factor influencing quality into an entity type and a property. (3) *Semantics*: The participants listed a number of concepts whose meaning was not clear to them. They did not completely understand the role of factors and quality requirements in the QMM. This list serves as input for improving the Quamoco QMM.

5.3 Feedback on the Evaluation Approach

Our objective was to evaluate of the Quamoco QMM, but also to test the developed evaluation approach for QMM generality, meaning (1) Is the developed approach applicable in practice? (2) Do we measure what we assume to measure? (3) Are the obtained results useful?

(1) The performed studies indicate the feasibility of the developed approach (no major problems occurred).

(2) One assumption is that the selected refinement of the construct generality is appropriate for describing it. A common way to check this assumption would be a

factor analysis. Due to the limited number of data points, such an analysis would not be reliable. To get a first impression of whether there is a relationship between the factors *completeness, understandability,* and *retained semantics* of the QM and the construct they should describe (generality of the QMM), we analyzed their correlation. Table 5 shows that completeness as well as understandability are positively correlated with generality (as we would expect). The retained semantics could not be checked due to a lack of variation in the obtained answers.

(2) Another assumption based on the chosen operationalization is that the measures defined for the different constructs are correlated with the perception of the participants and with their subjective evaluations. There is a highly significant positive correlation between *% of concepts supported* and perceived completeness and a highly negative correlation between *# unused and split concepts* and perceived understandability (Table 5). As previously mentioned, a correlation with the perceived retained semantics could not be checked due to a lack of variation in the answers.

Spearman's rho > 0.5: gray, p<0.10: bold	Perceived QMM generality	Perceived QM completeness	Perceived QM understandability	Perceived retained semantics
Perceived QMM generality	1.00			
Perceived QM completeness	0.63	1.00		
Perceived understandability	0.41	-0.11	1.00	
Perceived retained semantics	N/A	N/A	N/A	1.00
% of concepts supported	0.63	**1.00**	-0.11	N/A
# unused and split concepts	-0.39	0.00	-0.53	N/A
% of concepts explicitly sup.	0.40	0.42	0.24	N/A

Table 5: Correlation between perceived and measured constructs

(3) Finally, we want to know if generality is a QMM property from which the user of the QMM would benefit. In parallel to evaluating generality, we evaluated two constructs assumed to be affected by the QMM generality: *perceived usefulness* and *perceived ease of use* of the QMM. We assume that sufficient generality increases them. We measured these constructs with the *technology acceptance model* (TAM) questionnaire [5], a well-evaluated measurement instrument for these constructs. The TAM questionnaire states six questions with respect to the perceived usefulness of a technology and six questions with respect to its perceived ease of use. The subject answers each of these questions on a Likert agreement scale. Table 6 shows that ease of use is positively correlated with the perceived QMM generality, especially with the perceived understandability of the QM. The usefulness of the QMM is also positively correlated with generality, especially with the perceived QM completeness. Similar results were obtained for the collected measures (Table 7). Interpretation: The results show that the generality contributes to the perceived ease of use and usefulness of the QMM.

Spearman's rho > 0.5: gray, p<0.10: bold	Perceived QMM generality	Perceived QM completeness	Perceived QM understandability	Perceived retained semantics
Ease of use	0.61	0.61	**0.89**	N/A
learning_easy	0.73	0.73	0.37	N/A
model_to_do_what_I_want	**1.00**	**1.00**	0.36	N/A
interaction_understandable	0.61	0.61	**0.89**	N/A
flexible_to_interact	**1.00**	**1.00**	0.36	N/A
easy_to_become_skillful	0.79	0.79	**0.80**	N/A
easy_to_use	0.40	0.40	**0.92**	N/A
Usefulness	**1.00**	**1.00**	0.82	N/A
accomplish_tasks_m_quickly	0.56	0.56	**0.89**	N/A
improve_performance	0.61	0.61	**0.89**	N/A
increase_productivity	0.25	0.25	0.18	N/A
effectivness_on_job	0.82	0.82	0.83	N/A
easier_to_do_job	0.75	0.75	0.70	N/A
useful	N/A	N/A	N/A	N/A

Table 6: Correlation between perceived generality and TAM questionnaire results

Spearman's rho > 0.5: gray, p<0.10: bold	% concepts supported	# unused and split concepts	% concepts explicitly sup.
Ease of use	0.61	0.00	0.30
learning_easy	0.73	0.36	-0.13
model_to_do_what_I_want	**1.00**	-0.35	0.36
interaction_understandable	0.61	0.00	0.30
flexible_to_interact	**1.00**	-0.35	0.36
easy_to_become_skillful	0.79	-0.11	0.34
easy_to_use	0.40	0.11	0.23
Usefulness	0.78	-0.26	0.74
accomplish_tasks_m_quickly	0.56	-0.32	0.65
improve_performance	0.61	0.00	0.30
increase_productivity	0.25	0.00	0.54
effectivness_on_job	0.82	-0.74	0.78
easier_to_do_job	0.75	-0.58	**0.81**
useful	N/A	N/A	N/A

Table 7: Correlation between measured constructs and TAM questionnaire results

5.4 Threats to Validity

Internal validity / Selection of subjects. The difference in experience with the QMM between the subjects might have influenced the study results because objectivity regarding the QMM was not guaranteed. The small number of participants may have negatively influenced the study analysis because some tests could not be executed. Finally, the participation of the subjects in the project may have led to personal agendas that distort the results.

External validity / Maturity. Some participants did not answer all questions. These missing values could not be included for analyzing the data.

Construct validity / Definition of construct. The results show that understandability and completeness positively correlate with the perceived generality. A correlation with the perceived retained semantics could not be checked due to a lack of variation in the obtained answers.

Conclusion validity / Experimental design. The studies were conceived and executed as industrial case studies. No control of influencing factors was possible. The number of cases is limited and allows only limited validity of the conclusions. We mitigated this threat by using extremely different QMs from diverse companies.

6 Conclusions

We propose a generality evaluation approach for quality meta-models that primarily analyzes how appropriately a specific meta-model can express existing quality models. The approach also helps to investigate the perceived completeness, understandability, and usefulness of the models built with the meta-model.

We applied the evaluation approach to analyze the quality meta-model developed in the research project Quamoco and to transcribe six real-world, industrial QMs to this meta-model. Our study shows the applicability of the approach to realistic cases as well as the good correspondence of the measured features with the subjective rating of the people who transferred the models. The majority of the study participants considered the evaluated Quamoco QMM as sufficiently general. However, improvement potentials could also be identified.

For future work, we plan to analyze further QMs and standards and re-evaluate the Quamoco meta-model after improvements. The former is important, as the Quamoco meta-model aims to be a unifying meta-model for a broad area of QMs. Its generality and expressiveness need to fit this range of models. The latter stems from the iterative approach of Quamoco, which allows incorporating the feedback from the evaluation approach into the next version of the meta-model.

Moreover, we believe that the presented evaluation approach is not only applicable to quality meta-models, but also for evaluating the generality of meta-models in general. To show that, we also plan to apply the approach to meta-models that target domains other than software quality.

Acknowledgments

We thank all study participants for their contributions and Sonnhild Namingha for reviewing the paper. This work is partially funded by the BMBF project Quamoco (01 IS 08 023 B/C).

References

1. Bajaj A. The effect of the number of concepts on the readability of schemas: an empirical study with data models. In: Requirements Engineering, 9(4): 261-270, 2004.
2. Basili V and Weiss D. A Methodology for Collecting Valid Software Engineering Data. IEEE Transactions on Software Engineering, 10(3): 728-738, 1984.
3. Boehm BW. Characteristics of software quality: North-Holland, 1978.
4. Chen Y, Dios R, Mili A, Wu L and Wang K. An empirical study of programming language trends. IEEE Softw., 22(3): 72-78, 2005.
5. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quarterly, 13(3): 319-340, 1989.
6. Deissenboeck F, Wagner S, Pizka M, Teuchert S, and Girard JF. An activity-based quality model for maintainability. In: Proc. of the 23rd International Conference on Software Maintenance (ICSM 2007): 184-193, 2007.
7. Dromey RG.. A model for software product quality. In: IEEE Transactions on Software Engineering. 21(2): 146-162, 1995.
8. Genero M, Manso E, Visaggio A, Canfora G, and Piattini M. Building measure-based prediction models for UML class diagram maintainability. In: Empirical Software Engineering, 12(5): 517-549, 2007.
9. Guizzardi G, Pires LF, and van Sinderen M. An ontology-based approach for evaluating the domain appropriateness and comprehensibility of modeling languages. In: Model Driven Engineering Languages and Systems, vol. 3713 of Lecture Notes in Computer Science: Springer Berlin/Heidelberg, 691-705, 2005.
10. ISO/IEC 9126-1:2001 International Standard, Software engineering – Product quality, Part 1: Quality model, 2001.
11. Kitchenham B, Linkman S, Pasquini A, and Nanni V. The squid approach to defining a quality model. In: Software Quality Control, 6(3): 211-233, 1997.
12. Kläs M, Heidrich J, Münch J, and Trendowicz A. CQML scheme: a classification scheme for comprehensive quality model landscapes. In: Proc. of 35th Euromicro Conference on Software Engineering and Advanced Applications: 243-250, 2009.
13. Marinescu R and Ratiu D. Quantifying the quality of object-oriented design: The factor-strategy model. In WCRE04: Proceedings of the 11th Working Conference on Reverse Engineering, 192-201, 2004.
14. McCall JA, Richards PK, and Walters GF. Factors in software quality. Concept and definitions of software quality: final technical report Springfield: National Technical Information Service (NTIS), Reportnr. RADC-TR-77-369 (I, II and III), 1977.
15. Opdahl AL and Henderson-Sellers B. Ontological evaluation of the UML using the Bunge-Wand-Weber model. Software and systems modeling, 1(1): 43-67, 2002.
16. Winter S, Wagner S and Deissenboeck F. A comprehensive model of usability. In: Proceedings of Engineering Interactive Systems, LNCS: Springer, 2007.
17. Wohlin C, Runeson P, Host M, Ohlsson MC, Regnell B, and Wesslen A. Experimentation in software engineering, an introduction: Kluwer, 2000.