Evaluating the Quality of Process Models: Empirical Analysis of a Quality Framework

Daniel L. Moody¹,², Guttorm Sindre¹
Terje Brasethvik¹, Arne Sølvberg¹
¹Department of Computer and Information Science, Norwegian University of Science and Technology (NTNU), Trondheim, Norway N-7491
{dmoody, guttors, brase, asolvber}@idi.ntnu.no
²School of Business Systems, Monash University, Melbourne, Australia 3800

Abstract. This paper conducts an empirical analysis of the quality framework proposed by Lindland et al (1994) for evaluating the quality of conceptual models. 194 participants were trained in the concepts of the quality framework, and then used it to evaluate models in an extended ER language. A randomised, double-blind design was used, and the results evaluated by a combination of quantitative and qualitative analysis. Finally, an analysis was conducted of the framework’s perceived ease of use, usefulness and likelihood of adoption in practice. The study provides strong support for the validity of the framework and suggests that it is likely to be adopted in practice, but raises questions about its reliability. The research findings provide clear direction for further refinements of the framework to improve its reliability.

1 Introduction

Empirical studies show that more than half the errors in IS development are due to inaccurate or incomplete requirements [14], [16]. This suggests that more effort should be spent to quality assure conceptual models, in order to catch requirements defects as soon as they occur, or to prevent them from occurring altogether [27]. However, there are few guidelines for evaluating the quality of conceptual models. As a result, the quality of conceptual models is almost entirely dependent on the competence of the analyst [12], [19].

There have been a number of empirical or conceptual studies published on the quality of data or information models (e.g., [3], [9], [19], [20], [21], [26]). However, the topic of process model quality has been far less studied. The GOM framework [2] was evaluated in a case study in [25]. But our study is quite different from that one. Another process modelling language is used, namely APM [4], whereas [25] used an in-house modelling language of the company where the study took place. Moreover, a different quality framework is tested, namely the semiotic framework suggested by Lindland et al. [15]. Finally, the research method is different (experiment vs. case study).
1.3 Objectives of this Paper

This paper conducts an empirical analysis of the quality framework proposed by Lindland et al. [15]. The framework is based on semiotic theory (the theory of signs), and has been developed for understanding quality of conceptual models in general. There are three levels of quality:

- syntactic quality, whose goal is syntactical correctness,
- semantic quality, whose goal is validity and completeness, and
- pragmatic quality, the model should be understood by the stakeholders.

In this study the framework was used more instrumentally than originally intended, namely for evaluating model quality. The broad research questions addressed are:

1. Does the framework provide a reliable and valid basis for evaluating the quality of process models?
2. Is the framework likely to be adopted in practice?

The second research question addresses an issue that is rarely addressed in IS design research, despite the fact that this is critical for research to have an impact on practice.

2 Research Methodology

The Method Evaluation Model [18] is a theoretical model for evaluating IS design methods, which incorporates both aspects of method “success”: actual efficacy and adoption in practice. It combines pragmatic justification, a theory for validating methodological knowledge [24] and the Technology Acceptance Model [7], a theoretical model for explaining and predicting user acceptance of information technology. The constructs of the model and the relationships between them are summarized in Figure 1. Intention to Use and Actual Usage are determined by perceptions, which are in turn determined by actual efficacy. In Figure 1, the three central constructs represent internal, psychological variables, while the other constructs are external behavioural variables, which can be objectively measured. This model has been tested empirically [18].

![Figure 1. The Method Evaluation Model](image-url)
The research design is summarized in Figure 2. This shows the experimental treatment, experimental tasks, materials and dependent variables. A randomized, double-blind design was used.

**Figure 2. Research Design**
Participants and experimental treatment. There were 194 participants in this study, all third year students at the Norwegian University of Science and Technology. CS or IS students have been used in most previous experimental studies of conceptual modeling (e.g. [3], [9], [21], [26]). To prepare the participants for the evaluation exercise, the paper describing the framework was on the course reading list, and besides they were given a 2 x 45 minute lecture about it.

Materials and experimental tasks. 20 different case descriptions were prepared for use in the experiment. This was done to minimise the chance of collaboration between participants. Each participant then had two weeks of calendar time available for making a model according to the description. This model was submitted through a web-based system. Then, three reviewers were assigned to each model, and each participant was to review three models, belonging to three different cases, none of these being the one that the participant modeled (to prevent the reviewer from being prejudiced that her/his model was the right solution). There was one calendar week for completing all 3 reviews, which were also submitted through the same web-based system. Finally, the students were requested to answer a post task survey, also submitted through the same web based system. The survey consisted of 16 closed questions and 3 open questions. The closed questions consisted of 6 items used to measure Perceived Ease of Use, 8 items to measure Perceived Usefulness and 3 items to measure Intention to Use. Each item was measured on a 5-point Likert scale, using the opposing-statements format. The order of the items was randomised to avoid monotonous responses.

2.1 Dependent Variables

We distinguish between two types of dependent variables:

- Performance based measures: How well did subjects perform the evaluation task?
- Perception based measures: How effective did subjects perceive the framework to be in performing the evaluation task?

These types of measures represent the difference between actual efficacy and perceived efficacy as defined in the Method Evaluation Model.
**Performance Based Variables.** Four dependent variables were used to evaluate performance on the evaluation task: Syntactic Quality, Semantic Quality, Pragmatic Quality, and Overall Quality. A total of 564 ratings of each type (syntactic, semantic, pragmatic, overall) were collected as part of the study (an average of three per model). These were used to assess the reliability and validity of the evaluation framework (actual efficacy). All four quality ratings were given on a 7 point Likert scale, from 1 (poor) to 7 (excellent).

**Perception Based Variables.** We define three perception-based dependent variables for evaluating the framework: Perceived Ease of Use, Perceived Usefulness, and Intention to Use.

**Perceived Ease of Use.** Perceived Ease of Use is defined as “the degree to which a person believes that using a particular method would be free of effort”. This construct was measured using five items on the post-task survey (Questions 1, 10, 11, 14 and 16).
• PEOU1: Overall, I found the quality framework easy to use
• PEOU2: I found the quality framework easy to learn
• PEOU3: I found it easy to use the quality framework to perform the evaluation task
• PEOU4: I think I would now be able to successfully apply the quality framework to evaluate process models in practice
• PEOU5: The concepts of the quality framework are clear and understandable

**Perceived Usefulness.** Perceived Usefulness is defined as “the degree to which a person believes that a particular representation method will be effective in achieving its intended objectives”. This construct is measured using eight items on the post-task survey (Questions 2, 5, 6, 7, 9, 12, 13 and 15).
• PU1: Using the quality framework reduced the effort to evaluate process models
• PU2: Using the quality framework gave a more comprehensive evaluation of the models
• PU3: The quality framework reduced the subjectivity of the review process
• PU4: The quality framework helped me to identify errors in the process models
• PU5: Overall, I found the framework to be useful
• PU6: Using the quality framework made the evaluation task easier than if I had to rely on my judgement alone
• PU7: I think the framework would be useful in evaluating process models in practice
• PU8: Use of the quality framework enabled me to evaluate the process models more quickly than if I did not have such a framework

**Intention to Use.** Intention to Use is defined as “the degree to which an individual intends to use a particular method”. This construct is measured using three items on the post-task survey (Questions 3, 4 and 8).
If I have to review a process model in the future (either as a customer or an analyst) I will use this quality framework.

If I am working in a company in the future, I would recommend that they use this framework to quality assure their process models.

I intend to use the framework if I have to review a process model in the future.

The theoretical model. The relationship between these theoretical constructs and their empirical indicators are shown in Figure 3. In the diagram:

- Circles indicate latent variables or theoretical constructs.
- Rectangles indicate observed variables (in this case individual survey items).
- Dotted lines indicate measurement relationships.
- Solid lines indicate causal relationships.

As shown in the diagram, Perceived Usefulness is determined by Perceived Ease of Use, while Intention to Use is jointly determined by Perceived Ease of Use and Perceived Usefulness.

3. Results and Discussion

3.1 Summary of Analyses

The results were evaluated by a combination of quantitative and qualitative analysis techniques. Each of these analyses addresses a sub-question of one of the research questions defined at the beginning of this paper:

1. Reliability analysis: How consistently were reviewers able to apply the framework?
2. Validity analysis: Are the quality categories necessary, sufficient and independent determinants of process model quality?
3. Weight estimation: What was the relative influence of each quality category in decisions about the overall quality of a model?
4. Interaction analysis: What are the relationships between the quality categories?
5. Defect pattern analysis: What was the frequency of errors in each quality category?
6. Task accuracy: How accurately were participants able to identify quality defects using the quality framework (Type I, Type II and Type III errors)?
7. Perceived ease of use: How easy did participants find the framework to use?
8. Perceived usefulness: How useful did participants find the framework?
9. Intention to use: Were the participants likely to use the framework in the future?

Analyses 1-6 evaluate the framework’s actual efficacy in performing the task. Analyses 7-9 evaluate the framework’s perceived efficacy. Analyses 1-4 are quantitative analyses, analyses 5-6 are qualitative analyses, while analyses 7-9 involve a combination of the two (both closed and open questions were included in the post-task survey). Mixing qualitative and quantitative research methods can lead to a more comprehensive understanding of a phenomenon [10].

### 3.2 Reliability Analysis

Reliability was evaluated by measuring the inter-rater reliability between different reviewers of the same model. As shown in Table 1, levels of reliability of around .6 were found for each quality category. This means that around 60% of the variation is systematic. While there is no definitive standard for reliability, alphas of 0.7 or above are considered to be acceptable in the literature [22]. The observed levels of reliability are clearly lower than acceptable, which suggests that the quality framework in its current form cannot be reliably applied in practice.

<table>
<thead>
<tr>
<th>CONSTRUCT</th>
<th>CRONBACH’S α</th>
</tr>
</thead>
<tbody>
<tr>
<td>Syntactic Quality</td>
<td>.6159</td>
</tr>
<tr>
<td>Semantic Quality</td>
<td>.5778</td>
</tr>
<tr>
<td>Pragmatic Quality</td>
<td>.5682</td>
</tr>
<tr>
<td>Overall Quality</td>
<td>.6091</td>
</tr>
</tbody>
</table>

### 3.3 Validity Analysis

To assess the validity of the quality framework, we need to address the following questions (Alexander, 1968):

- **Completeness (sufficiency):** Do the set of quality categories cover all aspects of quality of process models? Are there any aspects that have been left out?
- **Parsimony (necessity):** Are all the quality categories necessary for evaluating process models? Are they all relevant determinants of process model quality?
- **Independence:** Are the quality categories independent of each other (disjoint, non-overlapping, mutually exclusive or orthogonal)?
A regression analysis was carried out using the quality categories (syntactic, semantic, pragmatic) as independent (predictor) variables and overall quality as the dependent (predicted) variable (Figure 4). The results of the regression analysis were:

- Adjusted $r^2 = 0.93$
- Significance level ($p$) = 0.000***

![Figure 4. Regression Model for Validity of Quality Categories](image)

**Completeness (Sufficiency) of Quality Categories.** The $r^2$ statistic measures the proportion of the variance of the dependent variable explained or predicted by the independent variables. The adjusted $r^2$ should be used in preference to $r^2$ in multiple regression to allow for shrinkage [23]. In this case, the quality categories account for more than 93% of the variance in Overall Quality. This is an extremely high value for $r^2$, which suggests that there are no missing variables in the model. This provides strong evidence that the set of quality categories is complete. Ratings of overall quality are almost entirely explained by variations in ratings for the individual quality categories.

**Parsimony (Necessity) of Quality Categories.** In a multiple regression, separate t-statistics and significance levels are calculated for each independent variable. These measure the effect of each independent variable on the dependent variable, while controlling for all other independent variable(s). As shown in Table 2, all independent variables (quality categories) were found to have a highly significant effect on the dependent variable (overall quality). This provides strong evidence that the quality categories are all relevant determinants of the quality of a process model. One should however note that there are possible threats to the validity of this result, as will be discussed in section 4.
### Independence of Quality Categories
Collinearity analysis was also conducted as part of the regression analysis, to evaluate the independence of the quality categories. There was no evidence of multicollinearity between the quality categories. As shown in Table 2, Tolerance values are well above .2 for all quality categories, which suggests that they are all independent determinants of overall quality.

<table>
<thead>
<tr>
<th>QUALITY CATEGORY</th>
<th>T-STATISTIC</th>
<th>STATISTICAL SIGNIFICANCE</th>
<th>TOLERANCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Syntactic</td>
<td>21.79</td>
<td>.000***</td>
<td>.437</td>
</tr>
<tr>
<td>Semantic</td>
<td>17.58</td>
<td>.000***</td>
<td>.345</td>
</tr>
<tr>
<td>Pragmatic</td>
<td>21.57</td>
<td>.000***</td>
<td>.379</td>
</tr>
</tbody>
</table>

### 3.4 Weighting of Quality Categories
Regression analysis was used to measure the separate effects of each quality category on judgements of overall quality. The resulting regression equation was (O: overall, Sy: syntactic, Se: semantic, P: pragmatic, Q: quality):

\[
O \cdot Q = 0.33 \cdot Sy \cdot Q + 0.31 \cdot Se \cdot Q + 0.34 \cdot Pr \cdot Q + 0.07
\]

The regression coefficients represent the influence of each quality category on decisions about the overall quality of a model. The result suggests that the categories are highly symmetrical in their effect, which is a positive finding. According to Alexander [1], it is important that all design variables are approximately equal in size and scope.

### 3.5 Interaction Analysis
Bi-variate correlation analysis was conducted to test for interactions between the different quality categories. Strong positive correlations were found between all quality categories:
- Syntactic and Semantic Quality (r = .72, p = .000***)
- Syntactic and Pragmatic Quality (r = .69, p = .000***)
- Semantic and Pragmatic Quality (r = .76, p = .000***)
Correlations are non-directional, but our theoretical explanation is that they represent causal relationships (as shown in Figure 5): Improving syntactic quality will improve pragmatic quality as a model that is not syntactically correct will be difficult to interpret. Improving pragmatic quality will improve semantic quality, as a model that is difficult to interpret will be difficult to match against the real world. Improving syntactic quality will therefore also improve semantic quality.
Three alternative multiple regression models were developed to test this explanation (these represent all possible causal combinations of the variables):

- Model 1: Syntactic Quality + Semantic Quality → Pragmatic Quality
- Model 2: Syntactic Quality + Pragmatic Quality → Semantic Quality
- Model 3: Semantic Quality + Pragmatic Quality → Syntactic Quality

The results of the analysis are summarised in Table 3. While all regression models were statistically significant, Model 2 (our a priori theory) clearly provides the best fit (as evidenced by the higher variance explained).

<table>
<thead>
<tr>
<th>MODEL</th>
<th>ADJUSTED R²</th>
<th>SIGNIFICANCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.619</td>
<td>.000</td>
</tr>
<tr>
<td>2</td>
<td>.654</td>
<td>.000</td>
</tr>
<tr>
<td>3</td>
<td>.561</td>
<td>.000</td>
</tr>
</tbody>
</table>

### 3.6 Defect Pattern Analysis

This was done by an expert who investigating the models and reviews. It was too much work to look through 194 models and 582 reviews, so a random sample of 32 models / 96 reviews were picked. Many of the models were of quite poor quality, and counting the number of errors did not give interesting results. A more interesting count was the number of students who made errors of various kinds. The most frequent mistakes were (percentage of students who made it):

- **lacking flows** (50%), i.e., at least one flow missing. Most typically, there were missing connections between resources and tasks, or missing connections between the input/output ports of a super-task and its first/last sub-task in a decomposition. Such defects may be more an effect of sloppiness than misunderstanding of the process flow of the case to be modelled. However, for about 20% of the students, there were also missing or misdirected flows between activities, so that the essential process flow itself was incorrectly captured.

- **missing roles**, (50%), and **missing resources** (44%) – the APM language puts roles in the lower part of the task nodes, and these should then be linked to resources of
various kinds (persons, tools, documents) filling the roles. Many models lacked at least one or two roles or resources somewhere in the diagram.

- **wrongly specified decision points** (35%), either using the wrong logical connective (and/xor), or (more commonly) having a decision point in the model without specifying its connective.
- **lacking decision points** (30%), totally omitting a decision point that should have been in the model).
- **poor naming of tasks** (27%).
- **missing task**, (25%) – normally just one.
- **wrong order of tasks**, (19%).

The bulk of defects were due to omissions (something that should have been in the model, but was left out). This may partly be due to poor modelling abilities and poor motivation for some of the participants.

### 3.7 Task Accuracy

Task accuracy was measured by inspecting the same 20 sample models and 60 sample reviews as mentioned above, counting the:

- **Ratio of Type 1 errors** (false negatives or errors of omission): where defects exist in the model but are not identified by reviewers. On average, each review reported just 2.4 defects, whereas the expert reported an average of 6.6 defects per model. Hence, on average, 64% of the defects went unreported. 20% of the participants did not care to report any defects at all (yet not necessarily giving top scores for the models), so if these are disregarded, the Type I ratio decreases to 55%.

- **Ratio of Type II errors** (false positives or errors of commission): when defects are identified which are not defects at all. In this respect the students actually did quite well: 95% of the defects reported by the participants were really defects according to the expert, so only 5% were Type II errors.

- **Ratio of Type III errors** (classification errors): where defects are correctly identified but classified in the wrong category. Here, the students did not do too bad either, only 9% of the errors were wrongly categorised.

As with the model defects, most review defects were omissions, i.e., defects that should have been reported but were not. Again, poor motivation seems to have been one important problem. Given their acceptable performance with respect to Type II and III errors, the students seem to have been able to find defects and classify them in terms of the framework, but only a few took the time necessary to find all the defects in a the given models.
9.8 Likelihood of Adoption in Practice

Validation of the Measurement Instrument. To evaluate the results for Perceived Ease of Use, Perceived Usefulness and Intention to Use, it is necessary to evaluate the validity and reliability of the survey items used to measure them.

Construct Validity. Factor analysis was conducted on the items used in the post-task survey using the principal components extraction method and varimax rotation with Kaiser normalisation. The initial factor analysis resulted in four factors being extracted using the Kaiser criterion. Q5 (Perceived Ease of Use) and Q14 (Perceived Usefulness) formed a fourth factor, and were therefore eliminated from the analysis. Factor weights derived from the analysis were then used to estimate the value of the underlying theoretical constructs.

Item Reliability. Reliability analysis was conducted on the items used to measure Perceived Ease of Use, Perceived Usefulness and Intention to Use (excluding Q14 and Q15). As shown in Table 4, high levels of reliability were found for all constructs, with Cronbach’s alpha > .7 in all cases, which is considered acceptable [22].

<table>
<thead>
<tr>
<th>Construct</th>
<th>Cronbach’s α</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perceived Ease of Use</td>
<td>.8614</td>
</tr>
<tr>
<td>Perceived Usefulness</td>
<td>.7737</td>
</tr>
<tr>
<td>Intention to Use</td>
<td>.9036</td>
</tr>
</tbody>
</table>

Significance Testing. One-sample t-tests were conducted for each construct to see if they were significantly different to 3 (the “zero point” of the Likert scale use). Table 5 summarises the results of the significance testing. Overall, participants found the framework easy to use and useful, and intended to use it in the future.

<table>
<thead>
<tr>
<th>Construct</th>
<th>Mean (μ)</th>
<th>Standard Dev (σ)</th>
<th>Significance</th>
<th>Y/N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perc. Ease of Use</td>
<td>3.94</td>
<td>.764</td>
<td>.000</td>
<td>Yes</td>
</tr>
<tr>
<td>Perc. Usefulness</td>
<td>3.89</td>
<td>.597</td>
<td>.000</td>
<td>Yes</td>
</tr>
<tr>
<td>Intention to Use</td>
<td>3.34</td>
<td>.941</td>
<td>.019</td>
<td>Yes</td>
</tr>
</tbody>
</table>

3.9 Analysis of Linkages Between Constructs

In the Method Evaluation Model (Figure 1), a number of causal relationships are hypothesised:
• Perceived Ease of Use (PEU) → Perceived Usefulness (PU)
• Perceived Ease of Use (PEU) + Perceived Usefulness (PU) → Intention to Use (IU)

As part of this research study, we evaluated the validity of these relationships using regression analysis. The resulting regression equations were:

$$PU = .329 \times PEU + 2.59$$ (2)

$$IU = .23 \times PEU + 0.71 \times PU \times 0.031$$ (3)

Both regressions were found to be significant with $\alpha < .005$. This means that the causal relationship between Perceived Ease of Use and Perceived Usefulness was strongly confirmed. The regression coefficients, significance levels and correlation coefficients for each independent variable in Equation 3 are shown in Table 6 below. These measure the effect of each independent variable on the dependent variable, while controlling for the other independent variable.

**Table 6. Multiple Regression Results**

<table>
<thead>
<tr>
<th>INDEPENDENT VARIABLE</th>
<th>REGRESSION COEFFICIENT (b)</th>
<th>SIGNIFICANCE (p)</th>
<th>EFFECT SIZE @</th>
</tr>
</thead>
<tbody>
<tr>
<td>PEU</td>
<td>0.026</td>
<td>(.885)</td>
<td>.021</td>
</tr>
<tr>
<td>PU</td>
<td>0.841</td>
<td>.001***</td>
<td>.534***</td>
</tr>
</tbody>
</table>

The relationship between Perceived Usefulness and Intention to Use was found to be highly significant with $\alpha < .01$, while the relationship between Perceived Ease of Use and Intention to Use was not statistically significant ($\alpha < .05$). This is consistent with many of the studies of the Technology Acceptance Model, which have found that the relationship between Perceived Ease of Use and Intention to Use is not significant after controlling for the effects of Perceived Usefulness [5], [7].

### 4 Discussion and Conclusion

This paper has conducted an empirical analysis of the conceptual model quality framework proposed by Lindland et al. [15]. The validation approach combines both quantitative and qualitative research approaches. The quality framework was found to be valid, with its categories found to be necessary, sufficient and symmetrical. It was perceived to be both easy to use and useful in evaluating process models, and participants intended to use it in the future. The research findings provide clear directions for future research, refining the framework to address the limitations identified. Finally, the paper yields a further empirical test of the Method Evaluation Model, which showed the measurement instrument had a high level of reliability and validity and validated two of the three causal relationships in the model.

The major limitation of this study is the nature of sample population used. In general, the population for a research study should be representative of the population to which the researcher wishes to generalise results [6]. For most laboratory experiments involving undergraduate students, external validity is a significant problem, as observed in [8], [11], [17]. Many of our participants delivered work of
poor or mediocre quality. Nevertheless, the fairly low amount of Type II and III errors in review indicate that the students were, at large, able to find and classify defects when making an effort to do so.

We do not see this study as the end of the validation process but only the beginning. We got useful empirical data about the framework and pilot-tested the validation approach, but the results must be interpreted with caution. For instance, students with little experience as modellers and reviewers may have been tempted to score "overall quality" simply as an average of their more or less sloppy scores for syntactic, semantic, and pragmatic quality (or vice versa, fitting the value for the three quality categories to an already assigned value for overall quality), although they were instructed that overall quality should be considered separately from the other variables. Such "averaging" would highly inflate the validity of the quality framework. Similarly, both validity and reliability might be inflated by defensive reviewing (i.e., assigning scores of 3-5 rather than 1-2 or 6-7). Hence, the generalisability of the results to practice is questionable. For this reason, our goal is to undertake similar studies with experienced practitioners, and possibly with refined versions of the quality framework. Some refinement of the framework has already taken place [13], but the experiences from this study suggest the need for other refinements, like customizing the framework to the modelling language used.

References