# On Business Process Model Reviews

Alexander Grosskopf and Mathias Weske

Hasso Plattner Institute, University of Potsdam, Germany
{alexander.grosskopf,mathias.weske}@hpi.uni-potsdam.de
bpt.hpi.uni-potsdam.de

**Abstract.** In process reviews, domain experts validate the model against reality. In general, reviews are conducted in an iterative manner. Better reviews can build consensus faster and save iterations, i.e. time and money.
In an exploratory study, student clerks were asked to provide feedback to models from their domain. In this paper, we report on the study and the review performance. We explore typical issues raised in reviews and derive implications for practitioners and further studies. We identified education as the most influential factor on review performance in our sample set.

**Key words:** Process Modeling, Reviews, Performance, BPMN, t.BPM

## 1 Introduction

Visualized process models serve as a communication vehicle in business process management. Moreover, they become the blueprint for software implementations. On the path from the initial business process elicitation to software support, review cycles are required. Models are created once and get iterated several times. Iterations typically involve feedback cycles with domain experts. They have to ensure that the domain knowledge is properly represented in the model. Better review performance promises less iterations, which in turn translates to time and money saved on projects. But what can you expect from domain expert's reviews? How can you influence the performance of the reviewing task?

Empirical research on business process modeling has largely investigated the roles of models [1, 2] and modelers [3]. Condensed findings from empirical research even led to modeling guidelines [4]. Process reviews have not been addressed comparably.

We did a pre-study to assess the experiment setup for t.BPM [5]. It is a tangible toolkit to enable BPMN process modeling on a table. As part of this, university freshmen were introduced to BPMN and filled in a feedback test about a given model (adopted from [2]). It contained a graph with 14 tasks and five block structured exclusive and parallel sections. Activities were labelled with *A,B,C...* The students easily passed a test on understandability (also adopted from [2]). It turned out that all of them, had a strong formal background. Even though the freshmen were barely educated in process modeling, they could map the process

B. Mutschler, J. Recker, R. Wieringa, J. Ralyté, and P. Plebani (Eds.):
CAiSE 2010 Workshop ER-POIS, Hammamet, Tunisia, pp. 31-42, 2010.

semantics to known concepts from mathematics and physics, such as logical equations and circuit diagrams. This obviously influenced their performance. We concluded that, to get meaningful data about process reviews, a more realistic setup is needed.

In this paper, we first introduce our study design in Section 2. The data is evaluated and discussed in Section 3. Additional insights are drawn from investigating a related study in Section 4. We close the paper in Section 5 with a discussion of the findings and implications.

## 2 Study Design

### 2.1 Setup

The sample population, used in research studies, should be representatives of the population to which the researchers wish to generalize [6]. Thus, we wanted domain experts to provide feedback to domain specific processes. From interviews with BPM consultants we identified a typical scenario in which process consultants give workshops to elicit the domain knowledge, model the processes, and send them out as email attachment. Domain experts are asked to provide feedback. The model gets iterated. Part of the workshops with the consultant would be reserved to educate the participants about the goal of BPM and the notation used for process modeling.

To emulate best practices in the field, we designed the following exploratory study for subjects at the trade school in Potsdam. Students there are learners to become office or industrial clerks. They get practical training on the job and theoretical background for their profession at the trade school. As clerks, we consider them to be representatives of the population to be generalized on. We chose the domain processes *Moving to a new flat* and *Getting a new job*. The seventeen students (18-22 years) are considered to be domain experts, meaning they do know the context and can comment on the processes. Additionally, we designed a two page introduction into BPM and a one page modeling sample (topic: *Making Pasta*). The sample page contained a legend of the BPMN elements used. On that same page four pragmatical hints for to process modeling were provided. In particular, we suggested the balanced use of gateways, an eighty percent rule for relevance to set granularity, verb-object style activity labels as suggested by Mendling et al. [4] and a notational convention for conditions at gateways.

The introduction and sample sheet were designed to condition the subjects. They replace the guidance provided by the modeling experts in the workshops. The written form enforced the same type of treatment for all subjects. This was embedded in a larger experiment design to test the effect of t.BPM [5] on subjects. The hypotheses were that t.BPM modeling would yield positive effects on individuals, including more feedback in process model reviews. While the experiment result are to be published, this study explores partial data with focus on feedback performance. The experiment procedure is depicted in Figure 1.

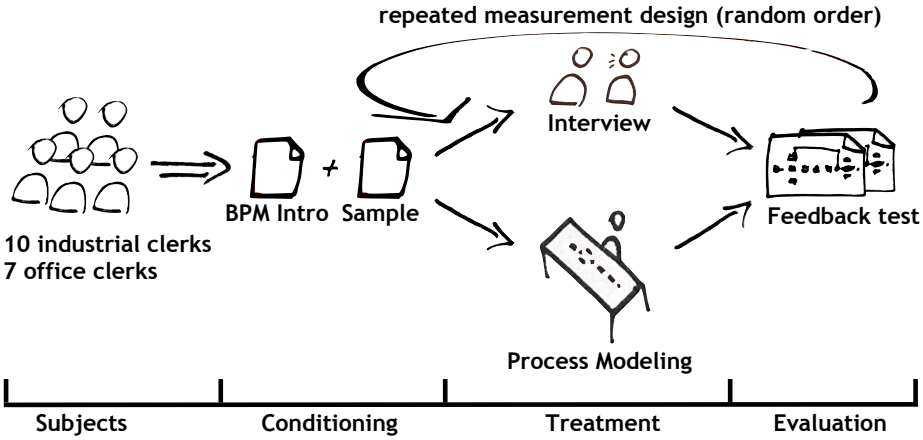



Fig. 1: Study Design

Each student got the BPM introduction and the sample. In general, time was not limited but tracked for each stage. Students were then randomly assigned to do either a structured interview or model their process on the table using BPMN elements. In that treatment step, they were asked to describe procurement processes, such as purchasing expensive hardware. Afterwards subjects were randomly given one of the process models for feedback. The treatment was repeated for each subject. In the second run, they got the alternative treatment and use the alternative feedback test.

In other words, the setup was a repeated measurement design in which all subjects get the same treatment in different orders. Subjects were assigned randomly. All subjects did interviews and process modeling. And all subjects did get both feedback tests, again randomly assigned.

### 2.2 Process Models used in Reviews

The process models used in the study are depicted in Figure 2. The models are annotated with the issues which we intentionally built into them.

Issues were chosen to belong to the area of *language* or *domain*. For example, two deadlocks were built into each process. This can be found by formal language analysis and requires no knowledge about the process domain. Nevertheless, we built-in these language related problems as indicators for the subjects' semantical understanding of the modeling language. The focus of this study are issues linked to the domain. They can only be interpreted if context information is available. Within the domain we consider three main categories: *labeling*, *information granularity* and *logical mismatch*. Labeling covers unsuitable naming of process elements, i.e. activity labeled with states not actions. Two obviously unsuited labels were build into the model (see Figure 2). Information granularity deals with too much or missing information in the process model. We left out an obvious activity and document per process model. Finally, logical mismatch describes wrong information in the model which contradicts the reality. We misplaced an

activity to generate an issue of this category. An overview of the built-in problems is given in Table 1 when we report on the review performance.

One sample issue, a missing control flow connector, was marked up in the model to indicate how to give feedback. We asked reviewers very broadly to "provide feedback". We assume that guiding questions and a clear focus, e.g. communicating the goal of the modeling effort, would have steered the reviews. Our goal is to explore. Therefore, neither guiding questions nor a goal were provided.

### 2.3 Variables

For this investigation *feedback* is the **dependent variable**. We quantify feedback by counting the number items provided in a review. Feedback is distinguished into intentionally built-in problems and additional comments. Categorization and quantification of feedback items was done by expert reviews. We refer to the sum of all issues raised as feedback. While the quality of feedback matters most, we start with quantity for our exploration. The initial assumption was that variation in the amount of feedback could be explained by the treatment method (t.BPM vs. interviews). Data analysis revealed no influence by treatment method (details in Section 3.4).

Thus, we decided to explore other available information to explain the variance in the data set. In Section 3.4 we investigate the *time*, the participant's *education*, and *sex* as **independent variables**. Guiding questions and modeling goal were consciously excluded as variables from this study.

## 3 Data Exploration

### 3.1 Data Analysis Instruments

The data in the sample set was tested and is normally distributed[1]. Significance was tested with a one-tailed t-test, abbreviated here with $p$. Correlation between variables was calculated using Pearson's correlation coefficient $r$. It is a normalized measure of dependence between two quantities where 0 indicates no correlation, -1 is a perfect negative correlation and 1 is a perfect positive correlation. In Section 3.4 we use Multi Regression Analysis to explain variation with significantly influential factors. The coefficient of determination $R^2$ describes the proportion of variation in the data set that can be explained with the regression model. As an example, $R^2 = .30$ means that thirty percent of the data variation can be explained by a particular regression model.

Based on the repeated measurements design we treat each test as an independent sample (n=34). We keep in mind that pairs of samples result from a single person, but we'll see that splitting them up yields no negative effect on the data analysis. In summary,

[1] True for Kolmogorov-Smirnov and Shapiro-Wilk test

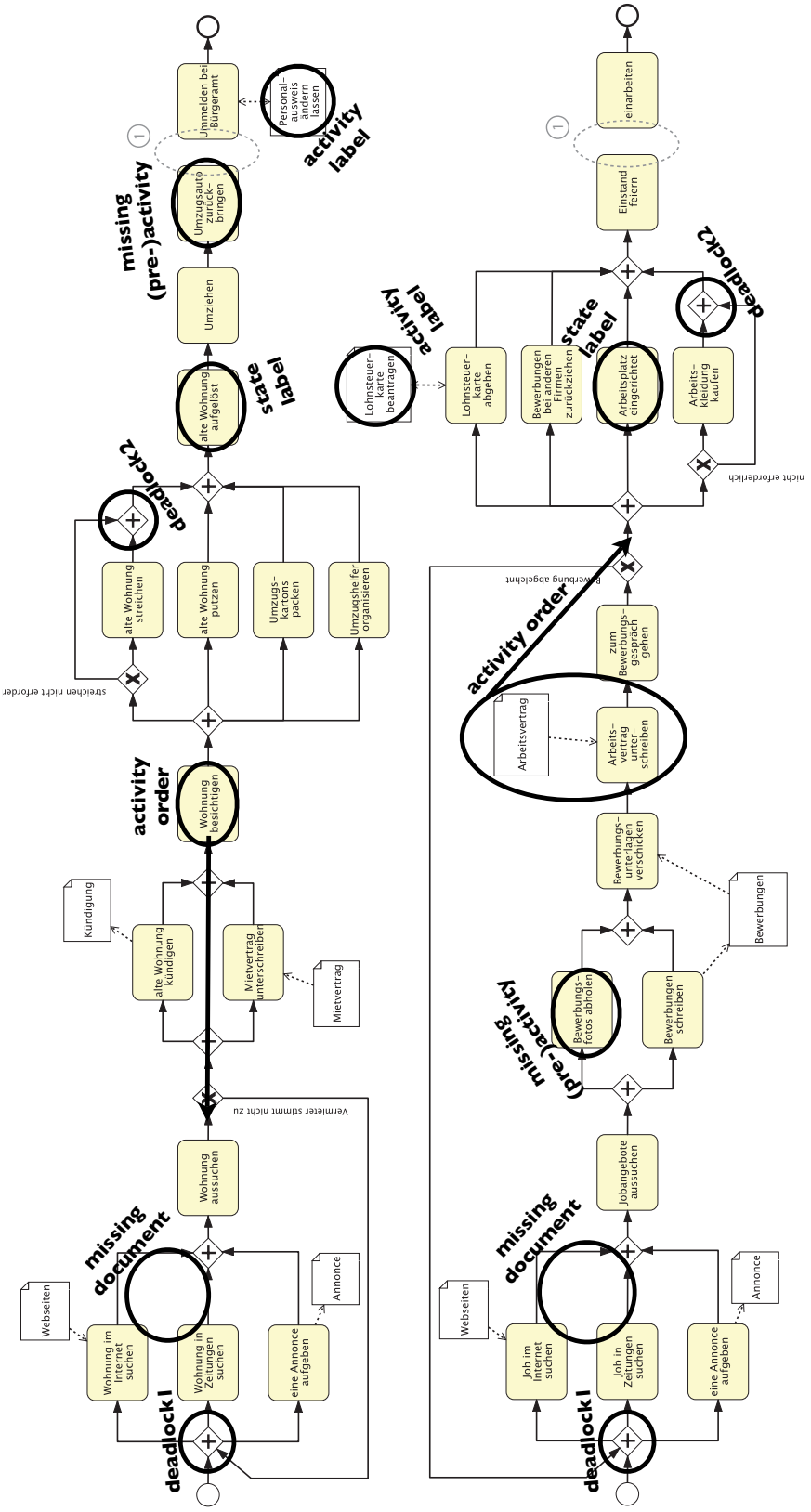



Fig. 2: Models used for feedback tests. Content is concerned with *Moving to a new flat* and *Getting a new job*

- data is normally distributed
- $r = [-1..1]$ describes the correlation of two quantities
- $p$ is a one-tailed t-test, $p < .05$ is considered as significant
- $R^2 = [0..1]$ is the explained variance in the regression model
- all numbers are based on a sample set of $n = 34$

### 3.2 Review Performance

The review performance of the subjects was quite poor. Most problems were not found with an overall success rate of less then thirty percent. Table 1 lists the built-in issues and shows how often it was by a single reviewer in one or both of the feedback tests.

While the issue of a *wrong activity order* was always found, by all reviewers in all feedback tests, the opposite is the case for the *deadlock1* which results from a loop back. If reviewers found an issue only once, it indicates that they did not systematically checked for this type of issue.

| **Category** | **Built-in Issues** | **Always Found** | **Found Once** | **Never Found** |
|---|---|---|---|---|
| Execution Semantics (Language) | deadlock1 (back loop) | 0 | 0 | **17** |
| | deadlock2 (bad block) | 1 | 3 | 13 |
| Labels (Domain) | activity labeled as state | 1 | 2 | 14 |
| | data object labeled as activity | 2 | 2 | 13 |
| Information Granularity (Domain) | missing document | 4 | 0 | 13 |
| | missing activity | 3 | 2 | 12 |
| Logical Mismatch (Domain) | wrong activity order | **17** | 0 | 0 |

Table 1: Built-in issues in the two review models, to be found by reviewers

Investigating the individual performance, we found that review performance varies between one and six reported built-in issues. On average, only two were found per person and indeed, in nineteen of the thirty four cases, only one built-in issue, the *wrong activity order*, was found. Reviewers gave 2.2 additional comments. In summary, over all tests, subjects reported back 4.2 feedback items on average. It shows that, although only a few problems were found (69 out of 238 in total), the reviewers still had a lot to share about the process with 75 items delivered as additional feedback.

### 3.3 Distribution of comments on topics

The role of comments is to capture additional issues which were not intentionally built-in. Domain specific issues with processes are not necessarily modeling mistakes, they might be a conscious decision to capture a certain aspect, or not. If an issue was arguable, it was counted as a comment. Almost all comments were counted, except for two. They were dropped as questions about the notation, not comments on the process. Comments were aggregated if they centered around one single issue.

| **Category** | **Addtionally Reported Issues** | **Amount** |
|---|---|---|
| Labels (Domain) | activity not labeled verb-object style* | 1 |
| Information Granularity (Domain) | missing document | 6 |
| | missing activity | 20 |
| | superfluous document | 1 |
| | superfluous activity | 10 |
| | missing event label* | 4 |
| | process scoping* | 1 |
| Logical Mismatch (Domain) | wrong activity order | 12 |
| | sequentialize parallel activities | 9 |
| | parallelize sequential activities | 2 |
| | lacking decision point | 7 |
| Ineffective Process | optimization potential | 2 |

Table 2: Additionally reported issues in 34 reviews

Simply counting comments was not meaningful, so we categorized them, see Table 2. Two researchers reviewed each comment, negotiated the type and category. Despite the potential experimenter bias, the advantage of this qualitative approach is to discover new issues and categories.

As shown in Table 2, most comments seek to inject additional information into the process model (30 of 75), rather than leaving them out (only 11).

Parallelizing or sequentializing activities was surprisingly popular (11 of 75 comments). As an example, subjects commented for the process "*Moving to a new flat*" that they would not clean until the are done with painting or that changing the address with the authorities should be started earlier in the process (in parallel). In our opinion it indicates, that the reviewers understood the semantics of the model as well as the domain. Two reviewers found fundamental optimization potential. As an example, if multiple offered flats were researched early on, we do not need to loop back and start over with research all the time. This observation is acknowledged by introducing a new issue and category. However, one might argue that this new category relates to process design (to-be situation) whereas feedback is typically focussed on validation (as-is situation).

Most interesting to note are the three **issues marked with *** in Table 2. Those three categories originate from four reviewers. Two of them criticized that start and end events were not properly labelled. One argued that the activity *"Moving"* should be a word-object style label. One reviewer raised the question for process scoping. In particular, he commented, that the process *Moving to a new flat* should be completed after the rental contract was signed. The subsequent activities are not in the scope of the process. While the authors do not agree with this opinion, it brings up process scoping as an issue addressed in reviews. Notes taken during the pre-study interview indicate that all four subjects were involved in process modeling activities within their company. We conclude, that they brought in additional process knowledge which was not part of the conditioning for this study. With nine out of eleven issue types being new, this qualitative assessment of feedback widened our or repertoire of issues addressed in process model reviews.

### 3.4 Influential factors

The initial assumption was that t.BPM modeling influences the reviewer's performance, which did not happen. Indeed, subjects performed quite stable in both feedback tests independent of treatment order or type, see Table 3 for details.

We even found that the amount of feedback does not significantly differ between the first and the second feedback test. For that reason we decided to treat all thirty-four feedbacks as independent samples ($n = 34$). We also compared the mean scores for the two different feedback models. They do not significantly differ which indicates that both models were equally hard or easy to understand. We therefore conclude that *model type, treatment and order have no influence* on the reviewer's performance.

*Sex, education, and time* taken to conduct the review had a significant influence on the performance. For education and sex the results are depicted in Table 3. Education emerges as the most dominant factor with the highest effect size and strongest significance.

| **Influence Factor** (independent variable) | **Alternatives** | **Effect Size** ($\overline{x}$ feedback) | **Significance** (one sided t-test) |
|---|---|---|---|
| Treatment order | 1st t.BPM / 2nd interview | 4.5556 / 4.7778 | .354 |
| | 1st interview / 2nd t.BPM | 3.875 / 3.75 | .418 |
| Treatment type | t.BPM / interview | 4.1765 / 4.3529 | .331 |
| Model type | moving / job finding | 4.0588 / 4.4706 | .15 |
| 1st/2nd Test | 1st / 2nd | 4.2353 / 4.2941 | .442 |
| Education | office / industrial | 2.50 / 5.45 | .000019 |
| Sex | male / female | 4.95 / 2.92 | .001 |

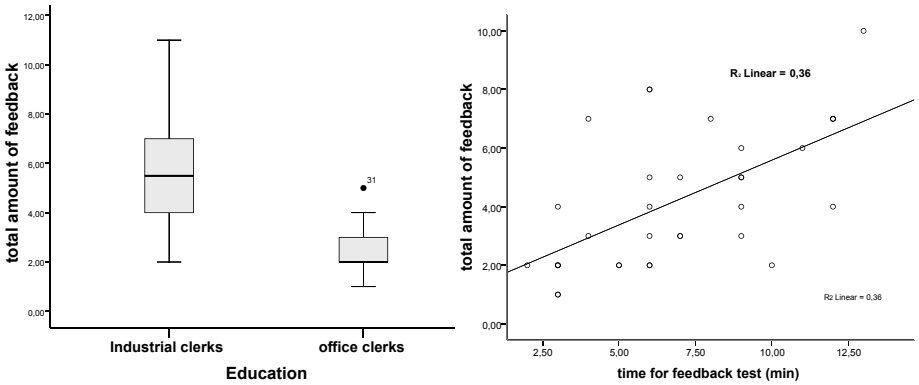
Table 3: Influential factors, individually tested for effect size and significance

To explain the significant ***influence of education***, we had post-study interviews with the principal of the trade school. We were informed that office clerks undergo a much stricter selection procedure and have better school achievements. On their job, they switch departments more easily and are often involved in supply chain optimization. Therefore, education for industrial clerks at the trade school does also include process notations, although in a very limited scope. Some students are also involved in process elicitation and modeling at their companies.

A boxplot in Figure 3a depicts the scattering of feedback for both groups. It visualizes that industrial clerks have much more to say about the process, up to eleven items in a single feedback test, while office clerks typically give two (at most five) items in a feedback test. In numbers, eight of fourteen tests done by office clerks reported one or two issues as feedback.

This dramatic difference due to education puts new light on our pre-study experience with HPI freshmen students. It raises the general question for transportability of empirical findings, if there is such a big gap between rather close professions.

***The influence of sex*** is likewise significant with a considerable effect size. However, there is also a large overlap of sex and education in our sample set. Out




(a) Boxplot comparing education: office clerks only provide 1-5 items with a median of 2

(b) Scatterplot and regression curve: The longer a feedback test takes, the more feedback is gathered

Fig. 3: The influence of education and time on feedback visualized

of eleven industrial clerks in the study, eight were male and three were female. Whereas out of six office clerks, two were male and four were female.

We conducted a hierarchical multi regression analysis to determine the actual influence of this variable. This multi regression model has a coefficient of determination of $R^2 = .508$, which means that it can explain 50.8 % of the variance in the data[2]. In that model, the contribution of sex boils down to explain 0,1% of the overall variance ($R^2 = .001$). The standardized multi regression equation is:

$$FEEDBACK = .404 * education + .346 * time_{feedback} - .093 * time_{intro} + .043 * sex$$

***The influence of time*** was determined using Pearson's correlation coefficient $r$. The time taken to complete the feedback test correlates significantly positive with the amount of feedback given ($p = .00008, r = .6$). That means, subjects that take more time for the feedback test, give more feedback. Figure 3b depicts the correlation in a scatterplot with a linear regression line. In the hierarchical multi regression model $time_{feedback}$ is the second strongest influence and contributes 10,4 % to the explanation of variance. While office clerks take about five minutes on average to complete the feedback test, industrial clerks take 8.3 minutes on average. We assume that subjects with less understanding have less to contribute and therefore need less time. Alternatively, subjects that investigate the process more deeply, find more issues but this of course needs more time. Similarly, people that need more time to read the BPM introduction perform worse in the feedback test ($p = 0.0375, r = -.39$). However, this has only a minor contribution of 0,9 % to the overall explanation.

***Concluding the review*** of the influential factors, we can explain 50.8% of the overall variation using a Hierarchical Multi Regression Analysis which considers the four variables $education$, $time_{feedback}$, $time_{intro}$, $sex$. The main influential factor is education with the highest significance and effect. Education alone can

[2] $R^2_{education} = .393129 \quad R^2_{time_{feedback}} = .104 \quad R^2_{time_{intro}} = .009216 \quad R^2_{sex} = .001225$

explain 39.3% of the data in the Multi Regression Model. The significance and effect size found for sex (see Figure 3), diminishes in the Multi Regression Model.

### 3.5 Limitations Discussion

The validity of this explorative study is limited by the decisions taken for its practical implementation. In particular, one might argue that the domain processes from a private background might limit the transportability of findings to business domains (external validity). And of course, the definition of an "issue" as well as its categorization is subjective (internal validity).

The small sample set, with the influence factors reported earlier, also limits the generalizability of findings. Larger sets with more controlled variables should be used for hypotheses testing. In this exploratory study, the small sample set enabled us to look deeply into the reviews (qualitative research). Thereby, we identified new issues that we did not see before.

Throughout the study and its evaluation we took the following countermeasures to limited the experimenter bias:

- We standardized conditioning for the subjects using written documents.
- Two researchers coded the feedback and negotiated categories.

## 4 Related Work on Reviews

In 2002, Moody et. al. assessed a quality framework for conceptual models using process modeling [7, 8]. The subjects were 194 third year students in Information Systems (IS) which had to model a process and then peer review three processes modeled by others. A set of 20 process models and their reviews was qualitatively investigated.

| **Category** | **defects** | **affected models** |
|---|---|---|
| Syntax (Language) | missing flows | 50% |
| | wrongly specified decision point | 35% |
| Labels (Domain) | poor naming of tasks | 27% |
| Information Granularity (Domain) | missing roles | 50% |
| | missing ressources | 44% |
| | missing activity | 25% |
| Logical Mismatch (Domain) | lacking decision point | 30% |
| | wrong activity order | 19% |

Table 4: Defects in process models created by third year IS-students — In peer reviews "*64% of the defects went unreported.*" [7]

The authors state that "Many of the models were of quite poor quality, and counting the number of errors did not give interesting results." [7]. Thus the "errors" were classified and the reviews were assessed. He uses the notion of defects to summarize the issues. Table 4 shows the defects.

Interestingly, subsequent expert reviews found 6.6 defects per model of which 2.4 got reported by the reviewers. In other words, "on average, 64% of the defects

went unreported" [7]. These numbers compare well with our seven intentionally built-in problems of which 2 were found (success rate $< 29\%$) on average.

While this is the nearest known relative to our study, several fundamental differences hamper a proper comparison of numbers from both studies. To name the most important ones,

- The review reported in Table 4 relates to modeling defects. In the study, reviews are evaluated by reporting true/false negatives/positives.
- The notion of defect used by [7] is much stronger than our notion of issues.
- Quality and defects per model did vary in [7], while we had a stable set of pre-defined issues per model.
- IS students have a very different education. They are method experts rather than domain experts.

Nevertheless, we learn from this study defect types that can be build into models for review tests. This further extends our set of feedback issues. Most important, we learn that proper education does not guarantee good process reviews. Thus, further research is needed to de-mystify the task of reviewing.

## 5 Discussion

Reviews by domain experts are a critical part of model validation and need more scientific investigation. Better review performance can avoid additional iterations needed in process analysis and design. This equals money and time saved on a project. We conducted an explorative study using qualitative and quantitative methods.

*Findings* from this study are the issue types and the influence factors. The identified issue types can be used to create better models for review tests with a larger variety of built-in issues. The distribution of issues raised by reviewers is also a finding. It can be used as a starting point to guide reviewers in their task. In other words, issues that are often missed might be worth a hint. Thus, reviewers can systematically check for them. A guideline for reviewers was out of scope for this work.

By statistical evaluation, we found education, sex and time as influential factors in the sample set. In particular, education dominated our findings. Although we had similar previous experiences with university freshmen, we did not anticipate education to be as influential within office and industrial clerks. We conclude that the subject group should be as homogeneous as possible to exclude those influences on the data set in future investigations. At the same time the model should involve a large variety of issues. Thus, it is possible to create the variance needed for insightful results. Our findings are limited by the small sample set and the dominance of education as an influential factor.

*Implications for practitioners* are phrased as suggestions to process modelers that do review cycles with domain experts. We suggest to,

- choose your reviewers wisely (huge differences in review performance)
- one reviewer per model is not enough (on avg. $> 60\%$ of issues not found)

*Further research* can built on the findings from this study to build a proper controlled experiment. In particular, the influential factors identified here should be fixed to rule them out. When designing models for review experiments, future research can take advantage of the domain related problems identified in this study. That can help to create models with a larger variety of problems built in.

In this study, we left out the aspects of a modeling goal and guiding review questions. We assume that they significantly influence the performance of reviewers. For example, guiding questions can link to frequently unreported issue types. In future work, we intend to investigate the influence of these aspects on reviewing performance.

## Acknowledgements

We gratefully acknowledge the support of Karin Telschow and Markus Guentert to setup and conduct the t.BPM experiment series. Special credits to Karin for her great support during the data exploration phase.

## References

1. Holschke, O., Rake, J., Levina, O.: Granularity as a Cognitive Factor in the Effectiveness of Business Process Model Reuse. In: Proceedings of the 7th International Conference on Business Process Management, Springer (2009) 260
2. Melcher, J., Mendling, J., Reijers, H., Seese, D.: On measuring the understandability of process models (experimental results). In: 1st Int. Workshop on Empirical Research in Business Process Management (ER-BPM). (2009)
3. Recker, J., Dreiling, A.: Does it matter which process modelling language we teach or use? In: 18th Australasian Conference on Information Systems. (2007) 356–366
4. Mendling, J., Reijers, H., van der Aalst, W.: Seven process modeling guidelines. Information and Software Technology (IST) (2009)
5. Grosskopf, A., Edelman, J., Weske, M.: Tangible business process modeling - methodology and experiment design. In Mutschler, B., Wieringa, R., Recker, J., eds.: 1st Int. Workshop on Empirical Research in Business Process Management (ER-BPM'09). (September 2009) 53–64
6. Cooper, D., Schindler, P.: Business Research Methods. 10 edn. McGraw-Hill Higher Education (2008)
7. Moody, D., Sindre, G., Brasethvik, T., Sølvberg, A.: Evaluating the quality of process models: empirical analysis of a quality framework. In: 21st Int. Conference on Conceptual Modeling–ER. (2002)
8. Sindre, G., Moody, D., Brasethvik, T., Solvberg, A.: Introducing peer review in an IS analysis course. Journal of Information Systems Education **14**(1) (2003) 101–120