# Students' Peer-Review in Modeling Exercises

***Guttorm Sindre***
***Norwegian University of Science and Technology and University of Auckland, New Zealand***
guttors@idi.ntnu.no

***Daniel Moody***
***Monash University, Australia and Charles Univ., Prague, Czech Republic***
dmoody@infotech.monash.edu.au

***Terje Brasethvik***
***Norwegian University of Science and Technology***

***Arne Sølvberg***
***Norwegian University of Science and Technology***

## Abstract

This paper reports on the 2002 experiences with students' peer-review of modeling exercises in a third year Information Systems course. While the peer-reviews did yield some positive and promising results, there were also some problems, and the goals set when introducing peer reviews were not fully achieved. Based on an analysis of student performance and perceptions, main causes for the problems were found to be lacking student motivation combined with unclear demands from teaching staff. For the 2003 offering of the course, the peer review process has been changed to remedy these problems, and this seems to have improved the learning experience for the students.

**Keywords:** information systems, conceptual modeling, inspection, peer-review

## Introduction

Peer-review or inspection (Fagan, 1976; Gilb and Graham, 1993; Wiegers, 2001) is a highly recommended practice in information systems development. It can remove defects more efficiently than testing (Gilb and Graham, 1993; Weller, 1993) and can also enhance learning and discipline in developer teams, reducing the number of defects introduced in the documentation in the first place (Gilb, 1999). Since there is more to gain the earlier defects are prevented or removed (Davis, 1993; Lauesen and Vinter, 2000), quality reviews should be strongly emphasized already in the early phases of IS development, such as system analysis and requirements engineering. But there has been significantly more research on code inspections, and in IS analysis courses there is much focus on modeling but less on reviewing the models.

Prior to 2002 there was little focus on reviews in the course Information Systems at the Norwegian University of Science and Technology. Then peer-review was introduced to improve the course. There were several gains that could be hoped for:

- Topical gains: Getting more focus on reviews as an important task in IS engineering.
- Pedagogical gains: There could be more feedback to the students without increasing the burden on teaching staff (Eschenbach, 2001; Thompson, 2001). While each teaching assistant had to look

at 25 models, each student would have to review only 3 models in a week, thus being able to give more feedback and hopefully learn from it at the same time.

- Staff workload reduction: The 2002 offering was partly a trial run to investigate whether peer grading (Gehringer, 2001) might be feasible in the future, as this might relieve staff resources for other activities, such as better course preparation or research.

The rest of this paper is structured as follows: Section 2 describes the course, and how it was changed to include peer-reviews in 2002. Section 3 presents the results of the evaluation of the peer-reviews both in terms of student performance and perceptions. Section 4 discusses the interpretation of the results. Section 5 then presents the changes that were made to the peer review exercises for the 2003 offering of the course, based on problems with the 2002 offering. Finally, section 6 concludes the paper, summarizing the results and suggesting further work.

## The Course

The course Information Systems has a reading lists composed of several book excerpts and journal papers. The core topic is conceptual modeling for information systems analysis, centered on a substantial excerpt from the book (Sølvberg and Kung, 1993). In addition to reading and attending lectures, the students are given exercises on a weekly or bi-weekly basis. Prior to 2002 the students delivered their completed exercises to teaching assistants, who would then correct them and make a pass/fail decision. With only 8 teaching assistants for about 200 students, each assistant would have 25 exercises to correct. Only hired for 100 man-hours for the entire course the teaching assistants could only provide superficial feedback to each student. Many students thus felt that they learnt little from the modeling exercises compared to the time spent.

For the 2002 offering, peer review was introduced in connection with the modeling exercises. The relevant exercises for the peer reviews were: 1) Process modeling in the APM language (Carlsen, 1997, 1998) – 2 weeks, *2) Review of process models – 1 week,* 3) Information modeling with the Referent language (Sølvberg, 1999; Brasethvik & Gulla, 2002) – 2 weeks, *4) Review of information models – 1 week*. Assignments 1 and 3 were done according to case descriptions in natural language, from 1-3 A4 pages. The review process resembled anonymous reviews of papers for conferences. Each student reviewed 3 different models, according to a semiotic framework for conceptual model quality (Lindland, Sindre et al., 1994). The review results were submitted through a web form. For each model, each reviewer provided:

- 1-7 Likert scale scores for the model's syntactic, semantic, pragmatic, and overall quality
- A list of identified defects in the model.

After each of the two reviews the students answered a questionnaire about the perceived usefulness of the semiotic framework for that task. At the end of the course they were asked to respond to a third questionnaire, investigating the perceived learning effectiveness of the peer review exercises. The two first questionnaires were compulsory to answer as part of the exercises themselves and thus had a 100% response rate. The third questionnaire was voluntary and had a 34% response rate. In the final exam, two of the questions required skills that should have been trained through the peer-review exercises, #1 to make an APM model (30%), #2 to review the quality of a use case description (20%).

## Results

The research results were obtained in several different manners: 1) Quantitative analysis of the students' performance, looking at the peer scores given for various model quality categories. 2) Qualitative analysis of the students' review performance, where an expert looked through 60 sample models and their corresponding reviews. 3) Quantitative analysis of the students' perceptions of the quality framework

that was used. This was done through a post-task questionnaire. 4) Quantitative analysis of the students' perceptions of the learning effects of the peer review exercises, as investigated through another questionnaire.

## Quantitative Analysis of Review Performance

The two main conclusions to be drawn from this were:

- The peer scores were not reliable enough to be recommendable for grading purposes. Most scores had inter-rater reliabilities around 0.6.
- The framework showed high validity (adjusted $r^2$ values around 0.9), i.e., multiple regression analysis confirmed a hypothesized causation from syntactic, semantic, and pragmatic quality to overall quality. The validity values may, however, have been inflated by the students' inability to make an independent decision on the overall quality of a model, as this should not simply be scored as an average of syntactic, semantic and pragmatic quality.

More details about the statistical analysis of these results can be found in (Moody, Sindre et al., 2002, 2003).

## Qualitative Analysis of Review Performance

The models to be reviewed were found to contain plenty of defects that could have been reported by the students. Yet many review reports were thin, some only giving quality scores for the models, not identifying any defects. Hence, the main qualitative observations were:

- About half the students did little work with the review exercises, reporting few defects. Yet the scores given were often quite average (typically 4's or 5's), while a model without defects should have been given perfect 7's. This indicates defensive reviewing. Other students used the more of the score scale. These students normally reported more defects, too. All in all, this means that scores would not made fair grades.
- Many defects went unreported, on average 60-70% for each model, although each model had 3 independent reviewers.
- Yet, when the students did report defects, they did this fairly well: About 90% of the reported defects were indeed defects (in expert opinion), and about 80% were even correctly classified according to the taxonomy syntactic, semantic, pragmatic. So it seems the students were able to find defects, but that they were too hasty or too poorly motivated to try to find many defects per model.
- There was much variation in the feedback received by various students. Some received 3 thin reviews. Others got feedback that it was possible to learn from. For the unlucky students, the peer reviews did not yield better feedback than previous years.
- Most disappointingly, the review performance deteriorated from exercise 2 to exercise 4 – a lower percentage of defects were reported, and more students delivered reports with empty defects sections for the latter exercise.

Some discussion of the quality of the review performance can be found in (Moody, Sindre et al. 2002, 2003; Sindre, Moody et al. 2003)

## Students' Perceptions of the Quality Framework

The responses to the post-task questionnaire show that most students did find the quality framework useful for the review task, scoring it at near 4 (of max 5) both for Ease of Use and Usefulness. However, it

must be noted that the students' limited experience makes it hard for them to answer questions about this. Since this was their first time reviewing ATM and Referent models, they would not know of any other way to do this than by means of the semiotic quality framework. So, it is difficult for them to know what the review might have been like with another framework, or with no framework. The student perceptions of the quality framework, as well as the instrument used to measure this, are discussed in more detail in (Moody, Sindre et al. 2002, 2003).

### *Analysis of Learning Effectiveness*

The responses to the end-of-course questionnaire about the learning effectiveness of the peer-reviews gave the following conclusions:

- Students felt that the course contributed positively to their learning, but the response was lukewarm rather than overwhelming. In particular, they were not convinced about the usefulness of the exercises with respect to long-term learning, i.e., beyond the course and its exam.
- The learning outcomes varied highly among the students, some claiming to have learnt a lot from the peer reviews, others claiming little or no learning.
- On average, the students spent less time than the nominally allocated for the peer reviews.

The questionnaire responses supports and partly explains the observations made in the qualitative analysis (poor reviews, defensive scoring, high variations in review performance). It is no surprise that performance is poor for students who spent little time on an activity, and then it is also quite understandable that they have learnt little from it. The design of the end-of-course questionnaire is discussed in detail in (Moody and Sindre 2003), and a more complete discussion of the learning outcomes for various student groups can be found in (Sindre, Moody et al. 2003).

## Discussion

Many students did poor work and learnt little from the peer reviews. We think the main explanations are:

- Poor motivation. This is supported both by the questionnaire responses and the qualitative analysis of review performance. It also came up in informal conversations with students.
- Unclear demands from teaching staff. It was possible to submit the review reports through the web system without identifying defects. This may have given the impression that only the scores were compulsory, while it was voluntary to list defects. De facto, this also turned out to be the case, as even those students who reported zero defects for imperfect models passed the review exercises. Hearing of this after the first review exercise, more students may have been tempted to deliver thin reports for the next review exercise, explaining the observed deterioration in performance.
- The chosen quality framework is on an abstract level, not customized to the particular modeling languages used. While this makes for a simple framework, which the students felt was easy to use (as confirmed by the post-task surveys), it also has some disadvantages in not giving the students detailed clues about how to review the models (as indicated by unreliable review scores and thin defect lists). In this respect, reviews based on more detailed checklists or quality standards might be better.

Partly the problems may have been caused by the fact that conflicts between the various goals of peer reviews, as presented in the Introduction, had not been resolved prior to the course. So, in a way, the peer review exercises fell between three different chairs:

- Anonymity (i.e., resemblance to double-blind conference paper reviews) was a chosen feature of the review process, to facilitate peer grading (Although scores were not used for grading in 2002, it was partly a trial run to investigate if this would be feasible). Without anonymity there would be a risk of cheating, students giving better scores to friends, etc. However, anonymity is not an advantage from a topical or pedagogical perspective. In industry, reviewers will normally know who has made a model, and possibly meet with the modeler to resolve misunderstandings. In pedagogy, studies in collaborative learning (Johnson and Johnson 1994) prescribe considerable face-to-face interaction, which clashes with anonymity.
- The lectures and reading material on reviews and the semiotic framework were selected according to topical concerns, and did not give sufficient guidelines on how to score a model (e.g., what defects to look for, what consequences various types and amounts of defects should have for the resulting score).
- A reviewer had little to gain from producing a detailed review. Even thin review reports passed, and since reviewer and reviewee were not cooperating towards a common goal, each student would be more interested in improving her/his own modeling skills rather than making a great effort helping unknown (anonymous) fellow students to improve their modeling skills.

Hence, in spite of anonymity, the resulting scores were far from reliable enough to be used for grading. Even if peer grading is the main goal for a peer review process, the students must be motivated, otherwise their work will be too poor to yield reliable grades. So our initial emphasis on total reviewer / reviewee anonymity throughout the whole process may have been unwise.

## 2003 Improvements

For the 2003 offering, we have attempted to improve the peer review exercises based on the evaluation of the 2002 experiences. The three most notable changes are:

- The review process has been shifted away from the anonymous "paper review" process of the 2002 offering. In 2003 there face-to-face meetings between reviewers and reviewees, and 4 students making models according to the same case descriptions are put together in a team, with a common goal.
- While in 2002 there was no separate activity for improving the models after review, this has been included as compulsory in 2003. This is more similar to industrial practice, and also more motivating, since the reviews are then needed to achieve the goal.
- The semiotic quality framework has been supplemented with some more detailed guidelines for how to review models of the concrete languages used. But it was chosen to retain the quality framework as overall guidance for the reviews, as the students responded positively to its Ease of Use and Usefulness for the task.

With these changes, the peer-review related exercises are now conducted in the following manner:

1. Modeling: Individually, the students receive a case description (20 different for the entire class, as in 2002) and make a business process model + an information model corresponding to this description.
2. Review: Four students who have all made models according to the same case description, reviewing each other's models. (In 2002, students had to review cases that were different from their own, which meant that they had to familiarize themselves with a new case per review. Now reviewing models for the same case, they are already familiar with the case description and can thus spend more time on the review itself – moreover, this gives them the opportunity to compare the models they review with each other and with their own solution to the same exercise).

3. Arena: The 4 students who reviewed each other now form a team and meet with a teaching assistant to discuss the reviews.
4. Modeling: Each student improves his model based on the review comments.
5. Review: Improved models are again reviewed within the team.
6. Arena: The four students meet with their teaching assistant again, discussing the new review results.
7. Modeling: Now the teams of four are going to select their best models and then make a joint extended model to be delivered in a report to be evaluated by teaching staff.

This new scheme clearly has some motivational advantages. Since the students who review each other are eventually going to deliver a joint report, it now pays off for one student to give thorough review feedback to another so that all the models within the team can be as good as possible (while in the 2002 run, the reviewer did not gain anything by helping the reviewee to improve). As stated in literature about collaborative learning, clearly perceived positive interdependence among the students is crucial (Johnson and Johnson 1994). A review process with several iterations to improve the same model is also more similar to industrial practice (in 2002, it was not compulsory to improve your model based on review comments), which can increase the perceived relevance of the exercises to future work-life.

## Conclusion

This paper has presented the experiences from the 2002 offering of an Information Systems course, where peer reviews were introduced in connection with two modeling exercises that the students performed. There were some positive experiences with the peer reviews, mainly:

- On average, the students reported that peer reviews did contribute positively to their learning, and qualitative investigations found that about half the students had worked quite a lot and probably learnt well from the reviews.
- The students perceived the semiotic quality framework used for the reviews as fairly easy to use and also as fairly effective for the review task.

But there were also negative experiences, in particular:

- About half the students worked little with the exercises and thus learnt little, and the feedback that these students produced would not have been valuable to the reviewees.
- The students were not convinced about the value of the review exercises for long-term learning (e.g., relevance to future work-life).
- The scores given in peer reviews were far from reliable. Luckily there was no intention to use them for peer grading that year anyway.

Major causes for the problems seem to have been poor motivation and unclear demands from teaching staff. Based on the experiences, some vital changes were made to the peer review process for the 2003 offering, as also presented in this paper. The new peer review process put more focus on motivating the students, both making the reviews more similar to industrial practice and organizing the students in teams with a common goal.

At the time of writing we are in the middle of the 2003 spring term, so the new review exercise scheme has not yet been evaluated, neither for review performance nor learning outcomes. But informal feedback from the students gives a much more enthusiastic impression than last year. Natural steps for further research will be to:

- Evaluate the students' review performance, both quantitatively and qualitatively, by the same means as for the 2002 offering. This will make it possible to compare results and see if their performance was indeed better after the changes.
- Evaluate the students' perceptions of their learning with the same end-of-course questionnaire as used in 2002, again to enable comparison to see if there has been any improvement.

Based on such evaluation, it will hopefully be possible to improve the peer review exercises even more. At the same time, other aspects of the course might also be evaluated (e.g., selection of reading materials, lectures, other exercises, exam question genres) to ensure that all attempted course improvements are consistent.

# References


Brasethvik, T. and J. A. Gulla (2002). A Conceptual Modeling Approach to Semantic Document Retrieval. *14th International Conference on Advanced Information Systems Engineering (CAiSE'02)*, Toronto, Springer Verlag.

Carlsen, S. (1997). *Conceptual Modelling and Composition of Flexible Workflow Models*. Doctoral Thesis, Department of Computer and Information Science, Norwegian University of Science and Technology.

Carlsen, S. (1998). Action Port Model: A Mixed Paradigm Conceptual Workflow Modeling Language. *Third IFCIS COnference on Cooperative Information Systems (CoopIS' 98)*, New York, IEEE Computer Society Press.

Davis, A. M. (1993). *Software Requirements*. Englewood Cliffs, NJ, Prentice-Hall.

Eschenbach, E. A. (2001). Improving technical writing via web-based peer-review of final reports. *31st ASEE/IEEE Frontiers in Education Conference*, Reno, NV, IEEE.

Fagan, M. E. (1976). Design and code inspections to reduce errors in program development. *IBM Systems Journal* **15**(3): 182-211.

Gehringer, E. F. (2001). Electronic peer review and peer grading in computer-science courses. *ACM SIGCSE 32nd Technical Symposium on Computer Science Education*, Charlotte, NC, ACM Press.

Gilb, T. (1999). Advanced Specification Inspections (Tutorial). *Quality Week '99*, San Jose, CA.

Gilb, T. and D. Graham (1993). *Software Inspection*. Reading, MA, Addison-Wesley.

Johnson, R. T. and D. W. Johnson (1994). An Overview of Cooperative Learning. *Creativity and Collaborative Learning*. J. Thousand, A. Villa and A. Nevin. Baltimore, Brookes Press.

Lauesen, S. and O. Vinter (2000). Preventing Requirement Defects. *Proceedings of the Sixth International Workshop on Requirements Engineering: Foundation for Software Quality (REFSQ'2000)*, Stockholm, Sweden.

Lindland, O. I., G. Sindre, et al. (1994). Understanding Quality In Conceptual Modelling. *IEEE Software* **11**(2): 42-49.

Moody, D. L. and G. Sindre (2003). The Leaming Effectiveness Survey (LES): An instrument for evaluating and improving the effectiveness of learning interventions. *Hawaii International Conference on Education*, Waikiki.

Moody, D. L., G. Sindre, et al. (2002). Evaluating the Quality of Process Models: Empirical Analysis of a Quality Framework. *21st International Conference on Conceptual Modeling (ER'2002)*, Tampere, Finland.

Moody, D. L., G. Sindre, et al. (2003). Evaluating the Quality of Information Models: Empirical Testing of a Conceptual Model Quality Framework. *25th International Conference on Software Engineering*, Portland, OR.

Sindre, G., D. L. Moody, et al. (2003). Introducing Peer Review in an IS Analysis Course. *Journal of Information Systems Education* (forthcoming).

Sølvberg, A. (1999). Data and what they refer to. *Conceptual Modeling, Current Issues and Future Directions (Selected Papers from the Symposium on Conceptual Modeling, Los Angeles, CA, held before ER'97)*. B. Thalheim. Berlin, Springer Verlag. 1565**:** 211-226.

Sølvberg, A. and D. C. Kung (1993). *Information Systems Engineering*, Springer Verlag.

Thompson, R. S. (2001). Relative validity of peer and self-evaluations in self-directed interdependent work teams. *31st ASEE/IEEE Frontiers in Education Conference*, Reno, NV, IEEE.

Weller, E. (1993). Lessons from Three Years of Inspection Data. *IEEE Software* **10**(5): 38-45.

Wiegers, K. E. (2001). *Peer-Reviews in Software: A Practical Guide*. Boston, Addison-Wesley.

## Bibliographies

**Guttorm Sindre** is an Associate Professor in the Dept of Computer and Information Science at the Norwegian University of Science and Technology (NTNU). He obtained a PhD in information systems from the NTNU in 1990, and his main research interests are conceptual modeling and requirements engineering.

**Daniel Moody** was a visiting professor at the NTNU at the time when this research was done. At the time of writing he holds a similar position in Prague (visiting from Monash University, Australia). He obtained a PhD in information systems from Melbourne University in 2000. His main research interests are data modeling and conceptual modeling. He is the current Australian President of the Data Management Association (DAMA) and Australian World-Wide Representative of the Information Resource Management Association (IRMA).

**Terje Brasethvik** is a PhD student at the NTNU. He received his MSc in information systems from the NTNU in 1996. Since 1998 he has been strongly involved in the teaching of the Information Systems course. His main research interests are information systems engineering and CSCW.

**Arne Sølvberg** is Professor in Information Systems at the NTNU (then NTH) since 1974, and is currently Dean of the Faculty of Information Technology, Mathematics and Electrical Engineering. He is the overall responsible for the Information Systems course. His main research interests are conceptual modeling and information systems development methods and tools. He has advised 20 PhD students over the last 10 years and been involved in several European research projects. He has been active in international organizations for research cooperation, such as the International Federation for Information Processing (IFIP) and the Very Large Databases (VLDB) Endowment, and is co-founder of the CAiSE conference series.