Introducing Peer Review in an IS Analysis Course

Guttorm Sindre
Daniel L. Moody
Terje Brasethvik
Arne Sølvberg
Department of Computer and Information Science
Norwegian University of Science and Technology
N-7491 Trondheim, Norway

ABSTRACT

Peer reviews were introduced as a teaching technique for the 2002 offering of the course SIF8035 Information Systems at the Norwegian University of Science and Technology (NTNU). The students handed in conceptual models which were then double-blind reviewed by three independent peers. The review reports contained scores and defect lists, but were not used for grading. This paper conducts an analysis of the outcomes of the peer review exercises. Several approaches are used, both quantitative and qualitative, investigating the students' performance and perceptions. The main conclusions are: 1) The scores given by the peer reviewers were not reliable enough to recommend their use for grading purposes. 2) The introduction of peer review exercises contributed positively to students' learning in the course – but not equally so for all students. A substantial fraction of the students did so little that it is hard to claim any learning effect. The main reasons for this seem to have been poor motivation and unclear demands. In hindsight, we have discussed the distinguishing properties of three different purposes of peer-reviews that were not clear to us when the course started, and we have identified several possible ways of improving the peer reviews. In spite of the reported problems, experiences are more positive than negative, and it has been decided to continue with peer reviews in the course.

Keywords: information systems analysis, conceptual modelling, peer-review, learning effectiveness

1. INTRODUCTION

In information systems development, it is a highly recommended practice to assure the quality of artefacts on all levels before they are used further on in the project. Such quality assurance activities, commonly known as inspections or reviews (Gilb and Graham 1993; Wiegers 2001), have been found to discover product defects far more effectively than testing of the developed code. The most effective discovery of defects occurs if inspections are applied systematically already at the requirements level (Kelly, Sherif et al. 1992). However, in the education of information systems engineers, review and inspection techniques have largely been neglected (Johansson 1997). This was also the case in the course SIF8035 Information Systems at the Norwegian University of Science and Technology, until the spring term of 2002. Then, a major change was made to the exercise part of the course, introducing peer-reviews of conceptual models. This paper constitutes a post-course analysis of this pedagogical intervention, trying to answer the following questions:

1. Would it have been reasonable to use the scores from peer-review for grading? (The peer scores had no impact on final grades in this offering of the course, but the 2002 offering could partly be considered a trial run to investigate whether peer grading would be a way to go.)
2. Did the introduction of peer reviews improve the students' learning in SIF8035 Information Systems?
3. Would it have been better if peer reviews had been introduced in another way?
4. Should peer reviews be included in IS Analysis and Design courses in general, and what teaching guidelines can be extracted from our experience?

These questions are listed in increasing levels of ambition. The first can be fairly well answered from our data, while the answers to the latter two will mostly be subjective opinions, although based on experiences from the review exercises. The rest of the paper is structured as follows: Section 2 discusses various purposes of peer-review in education. Section 3 presents the course as it was before peer-review was introduced. Section 4

describes the peer-review activities in more detail. Section 5 presents research methods and results. Section 6 discusses the interpretation of the results. Section 7 makes the conclusion of the article, whereupon section 8 discusses possibilities for further work.

2. DIFFERENT PURPOSES OF PEER REVIEWS

In education, peer-review would mean that students evaluate each other's work. (Newell 1998; Eschenbach 2001). Peer- review can be introduced for several different purposes:

- (Peer) Reviews may be a topic in the course. Peer is put in parentheses here, since in industrial IS development projects, reviewers may sometimes have quite different roles and backgrounds than those who wrote the documents submitted to review (e.g., a requirements specification may be written by analysts but reviewed by customer representatives).
- Peer-review may have positive pedagogical effects.
- Peer-review may reduce the work burden of the staff.

Of course, one may have all these three motivations at once. But there are some potential conflicts between them. As argued in section 1, review techniques may be a legitimate topic in an IS course, a learning goal then being to train the students to become good reviewers in later work-life. With this review-as-topic outlook one might expect to see that

- The review exercises are backed up by theory on review techniques, through lectures and reading materials.
- The suggested review process(es) resemble industrial review processes. While some differences may be necessary, the reviews performed at school should be relevant in preparing the students for industrial reviews.
- The reviews (process and/or reports) would be subject to assessment by teaching staff, as a source for mentoring.
- Review of a given work-product could be a potential exam assignment.

An early example of a university course with review-as-topic is reported in (Collofello 1987), using lectures, exercises, and a team project including several types of technical review.

With a pedagogical motivation for the review exercises, reviews need not be a topic in the course. An IS course could easily have modelling as a learning goal, not reviewing, and still include peer-reviews as a learning activity, with the assumption that modelling is better learnt by modelling + reviewing than by modelling + more modelling. For instance, reviewing fellow students' models, and receiving peer feedback to own models, could stimulate deeper thinking about modelling (cognitive benefits of peer reviews). Or the students might put more work into their models to make a good impression on their peers (motivational benefits of peer reviews). Then, in contrary to review-as-topic, the review-as-pedagogy outlook could imply that:

- There might not be theory about reviewing in the course, only about (e.g.) modelling.
- Rather than striving for industry-resemblance, the chosen review process would strive for pedagogical effectiveness, maximizing the learning outcome rather than the quality improvement of the reviewed artefacts.
- Teaching staff would primarily assess the models (or in general: the documents reviewed), not the peer-reviews. Teachers would perhaps inspect the delivered reviews only if the reviewee disagrees with the reviewer, seeking a second opinion.
- Reviewing would not be that likely as an exam assignment, as it was only a vehicle to learn something else (e.g., modelling), not a topic in its own right.

Industry resemblance may often be motivating, thus positive also to pedagogical effectiveness. But it is also easy to find examples where industry-resemblance and pedagogy would clash. For instance, when work of poor quality is submitted, a feasible industry response might be flat rejection. For instance, an IS requirements document may be found too immature too be feasibly reviewed in detail, instead simply sent back for rework. Similarly, an editor may abandon a manuscript of fiction only after a few pages, the only response to the author being a brief standard rejection letter. In an educational setting (e.g., a course in requirements engineering or creative writing), such flat rejection would hardly be acceptable. Students will have better chances of improving if given details about why their work was too poor, as well as encouragement about good aspects of the work, and even the less clever would have right of feedback, at least if making an honest attempt at the task.

One case where peer-review was used mainly for pedagogical purposes is (Eschenbach 2001), a course in technical writing. Here, reviewing was not a major learning goal, rather the point was that the students' writing skills would be more effectively trained if they peer-reviewed their reports.

Finally, work-reduction is the third possible motivation for using peer-reviews in a course. If students can comment on errors and suggest improvements on each other's exercises throughout the term, teaching staff time can be freed for other purposes, such as providing better reading materials, assignments and lectures. The highest savings of staff effort can probably be made if the peer-reviews yield scores that can be used for grading purposes, and especially if peer-grading can be conducted over the web (Gehringer 2001). Although peer-grading is commonly used in some educational contexts, there is limited knowledge about its validity

(Thompson 2001). With peer grading, there will be obvious issues of fairness and reliability of the grades, hence the review-as-work-reduction outlook would typically imply that

- There would be some theory on reviewing in the course. But unlike review-as-topic, this would not be theory about industrial review techniques, but rather theory about grading (e.g., meaning of various grades in terms of work product qualities, typical mistakes to look for, how many points to deduct, etc.)
- For the review process, one would want *anonymity* between reviewers and reviewees, and *independence* of reviewers. Without anonymity, there is a danger that students would give better grades to friends. The use of several independent reviewers for the same work product may improve grade validity (e.g., if one student grades inadequately, there will be several others to average it out). The use of several independent reviewers is also found in some industrial review techniques, and could be pedagogically motivated (e.g., to prevent lazy students from copying reviews off others). But anonymity is uncommon in industrial reviews. Neither is it likely to be considered a pedagogical advantage, since face-to-face discussion might enhance both topical learning and communication skills.
- The reviews might be subject to teacher assessment. But unlike review-as-topic, the primary objective would not be to help the students improve as reviewers, but to check whether grades are fair. Such checks must require significantly less time than it would have taken the teacher to grade the work in the first place. Hence, the teacher can only give all delivered assignments and reviews a superficial look, or possibly take samples for more thorough consideration (e.g., re-evaluate only those grades that have been disputed).
- Reviewing would not be a likely exam assignment.

The above discussions show that there are potential conflicts between review-as-topic, review-as-pedagogy, and review-as-work-reduction. Hence, it is not trivial how to implement peer reviews in a course, especially if the teacher has more than one of the three motivations. As will be seen further on in this paper, the various motivations and possible conflicts between them had not been sufficiently sorted out before the 2002 offering of SIF8035, so much of the above is hindsight developed on basis of problems experienced.

3. THE COURSE BEFORE PEER REVIEW

SIF8035 Information Systems is taught every spring term at the Norwegian University of Science and Technology (NTNU). The course gives 2,5 credits (Norwegian: *vekttall*) and is one of four equal-size courses that the full-time student will take that term. There are currently approximately 200 students taking the course at each offering. Most of these (>150) are third year students in a five years Masters program in information technology. For these, the course is compulsory within the degree. The remaining students are mostly from other related Masters programs (business administration, electrical engineering). These take the course on a voluntary basis. The prerequisite for the course is SIF8018 Software Engineering, which the students take as compulsory in their second year. SIF8018 currently uses the textbook by van Vliet (van Vliet 2000).

The curriculum of SIF8035 itself covers the following topics: information systems analysis and design, software requirements engineering, Enterprise Resource Planning systems, and human computer interaction. The uniting theme of these various topics is modelling, hence the course can be looked upon as a tour of various modelling approaches (like Data Flow Diagrams and related languages for business process modelling and task modelling in user interfaces, Petri nets and formal modelling languages built on logic, Entity Relationship diagrams and related languages for semantic information modelling, object oriented class diagrams, and use cases). It has been difficult to find one textbook covering these topics in an appropriate manner, so the course materials are composed of various book and journal excerpts. There are substantial inclusions (more than 40 pages) from the following books: (Sølvberg and Kung 1993) – 194 pages, (Preece, Rogers et al. 1999) – 150 pages, (Bancroft, Seip et al. 1997) – 61 pages, (Kotonya and Sommerville 1997) – 43 pages, (Kulak and Guiney 2000) – 43 pages. In addition, the reading list includes smaller excerpts from the books (Fowler and Scott 2000), (Farschchian 2001), (Scheer 1998), (Scheer 1999), (Curran and Ladd 2000), and the papers (Lindland, Sindre et al. 1994), (Sølvberg 2000), (Parsons and Wand 1997), (Davis 1995), (Sawyer 2000), (Maiden and Ncube 1998), (Hirschheim and Klein 1989), (van der Aalst 1999), (Kirchmer 1999), (Gulla and Brasethvik 2000).

Prior to 2002, the following teaching methods were used:

- Plenary lectures covering reading list material, normally 4 x 45 minutes a week throughout the 14 week term.
- Compulsory exercises, evaluated by teaching assistants as pass/fail. Failed exercises (unless redone) would mean loss of right to sit the exam. The most frequent exercise tasks have been to make models / specifications according to natural language case descriptions. One or two weeks are given for the completion of each exercise. Before peer-review was introduced, the standard procedure would be that students submitted their completed exercises to teaching assistants. The teaching assistants then evaluated them by a pass/fail decision, possibly giving comments about weaknesses in the student's answer. Since each

teaching assistant is responsible for a significant number of students, the feedback would normally be fairly superficial.

- The exam itself (5 hours written, no books). At the NTNU, exam questions of previous years and suggested answers are normally available to students, who often look to these for ideas of how the next exam will be. Hence, the learning effect of previous exams cannot be underestimated (for better and worse). The questions are different from year to year, but the students will have certain expectations of question topics and genres and often prepare tactically according to this.

Some might wonder why the course does not have a bigger project instead of small exercises. But there is indeed a project connected to the course, namely SIF8080 Customer Driven Project (Andersen, Conradi et al. 1994). This is taken by the IT students in their fourth year and has 5 credits, twice the weight of SIF8035. So, the students learn the necessary theory in SIF8035, then practice it in SIF8080 the following term. In SIF8080, each team is given a real customer with a real problem to be analysed and solved. Thus the project in SIF8080 is much more realistic than what could be achieved within SIF8035, with only 2,5 credits that would mostly have to be focussed on theory.
Whereas the project course SIF8080 has been positively evaluated both by students and alumni (Sorge 2000), it seems that many students are less motivated for the preceding theory course SIF8035, and the exam results have been somewhat disappointing. Peer-review was introduced in 2002 to improve the course.

4. TEACHING AND RESEARCH METHODS

4.1 Overall discussion of methods

Peer-reviews were introduced in the course in the spring term of 2002. The review activities were part of the mainstream teaching of the course, applying to all 200 students taking it. Hence, part of what is discussed under the heading research methods here could just as well have been called teaching methods. Anyway, it makes sense to present teaching and research methods as two sides of the same coin, since data were collected from the learning activities, whose outcome was also the focus for research.

There are some notable challenges to our ability to answer the research questions:

- We had no "control group" for the research, i.e., a group of students not participating in the peer review exercises.
- The student activities were not performed in a controlled setting. Students were free to do the exercises in the lab or at home, and were not supervised during the performance (except that, in the lab, they were free to ask for advice from teaching assistants). Hence, it might be possible for students to collaborate, and the time they spent on the exercises may have varied a lot.

As part of the mainstream teaching of a big course the exercises had to be performed in a way that was administratively feasible and in accordance with normal university practice. The absence of a control group can also be motivated by some fundamental problems in assessing hypothesized pedagogical improvements. Ideally, one would think that a comparison of student performance would be the right way to go. One group of students would perform peer review exercises, the other group would not, and then one could compare their performance, e.g., at the final exam. However, this comparison is difficult to set up. The two possibilities would be between-year comparisons (e.g., comparing students of 2001, who did not do peer reviews, with students of 2002) and within-year comparisons (e.g., dividing the 2002 class in two groups, one doing peer reviews and the other not). There are fundamental problems with both these approaches. The between-years comparison only allows for quasi-experimental designs because the performances are not totally comparable. The students are different, the exam questions must be different, and it is impossible to eliminate other changes in the learning situation that might affect student performance, such as teacher motivation, lecture times, exam scheduling, and changes in the workload of other parallel courses. The within-year comparison can be a true experimental design, but there are serious challenges. It is difficult to randomly divide the class in two equally clever groups and isolate them from each other. So, there will be a diffusion of treatments, e.g., through discussions and borrowing of notes. Even worse, it would be considered unethical to give different treatments, especially when one is believed to be an improvement over the other. Many universities have a policy against experiments that might affect student grades. Finally, a challenge to both approaches is that even if exams were comparable, one cannot be sure that a measured improvement in exam performance has not come at the cost of reduced knowledge in parts of the course curriculum not addressed in that year's exam questions.

Our choice is thus to look only at the students of 2002, all receiving the "treatment" of peer review exercises. Then, it is hard to draw strong conclusions from their performance alone, since one cannot know how their learning would have been if the peer reviews had been replaced with something else. Instead, we will use triangulation of method (Jick 1979), supplementing quantitative and qualitative investigations of student performance with an investigation of their perceptions, i.e., how they evaluate their learning experience. It may be contended that students lack the knowledge and experience to assess the quality or importance of their learning, but various studies show high correlations between students and more experienced groups (e.g., alumni, university professors, university administration) in evaluating teaching (Felder 1992). Hence, looking at

both performance and perceptions and applying both quantitative and qualitative methods, it should be possible for us to answer our questions more fully.

4.2 Activities Undertaken By The Students

Below we list the activities regarding the peer reviews. The only activity that the students had to conduct purely for research purposes was the answering of a questionnaire (item 10). All other activities were part of the teaching as such. Items in square brackets are activities that would have been undertaken also without the introduction of peer reviews (as in the 2001 offering). These have been included for comprehensiveness, as it is hard to conduct peer reviews without first producing something that can be reviewed.

1. [The students received lectures about process modelling in the APM language (Carlsen 1997; Carlsen 1998).]
2. [The students received, as a lab exercise, a tutorial of the METIS™ modelling tool (Computas 2001), that would be used in the next exercises.]
3. [Each student accessed his/her case description through the web system specifically designed to administrate the exercises. Virtually, this system made it seem that there were as many different case descriptions as there were students, i.e. student #1 had case #1, ..., student #200 had case #200. In reality, there were 20 different case descriptions, which is still a lot more than previous years, when all students used the same case description. Each case was written in natural language (Norwegian), ranging from 1-3 pages A4, and describing some company with a claimed information processing need.]
4. [Each student individually made a business process model in the APM language, based on the case description assigned to that student. Two calendar weeks were available for this task, but the nominal workload only 8 hours (4 per week). This is because the students take 4 courses in parallel each term, so only 12 hours per week belonged to the Information Systems course, of which 8 hours per week were stipulated for attending lectures and reading compendium material.]
5. The students received lectures relevant to the review task, such as theory on the semiotic framework for conceptual model quality (Lindland, Sindre et al. 1994), according to which the models should be evaluated, and a demo of the web system through which their reviews had to be submitted.
6. The student performed reviews of APM models made by their peers. The allocation of models to reviewers was made automatically, according to the following criteria:
   - Each student would receive 3 different models to review, none of which were based on the same case description that the student had modelled in point 4.
   - Each model would receive 3 different reviews.
   - Reviewers would not know the identity of the modeller and vice versa (i.e., double-blind design)
   - Neither would reviewers know the identity of other reviewers evaluating the same model.

   To review a model, the student would have to look at the model and the corresponding case description, and then:
   - Assign four scores to the model, all Likert scale ranging from 1 (poor) to 7 (excellent). The scores were for syntactic quality (i.e., conformance with the grammar rules of the APM modelling language), semantic quality (i.e., conformance with the information given in the case description), pragmatic quality (i.e., how easily understandable was the model), and overall quality (in the reviewer's own opinion, independently of the three former scores).
   - Describe defects found in the model, in free text. The defects were to be listed and categorised as syntactic, semantic or pragmatic defects.
   - Make general comments about the model in a separate free text field. Here, it would also be possible to include report of defects (if any) that the reviewer felt did not fit into the categories syntactic / semantic / pragmatic.

   The results of a review were to be submitted through a web form. The way the form was designed, it was technically impossible to assign anything but scores within the 1-7 range, and it was also impossible to submit it without scores. But it was possible to submit the form without reporting defects or giving general comments. From the submitted reviews, the system automatically generated review reports. These could be accessed through the web system by modellers, teaching staff, and researchers.
7. [The students received lectures about information modelling in the Referent language (Sølvberg 1999; Brasethvik and Gulla 2002).]
8. [The students individually made information models in the Referent language. These were based on the same case descriptions as used in task 4, and with the same time frame and nominal workload (2 calendar weeks, 8 hours in total).]
9. The students peer-reviewed the information models. This was done in the same way as with the process models (task 6), with the same time frame and nominal workload (1 calendar week, 4 hours), and reported through the same web system.
10. In connection with the last lecture before the exam, a questionnaire was distributed among the students, investigating how they perceived the learning outcomes of the peer review activities. The questionnaire was developed specifically for this purpose. The students had 10 minutes to complete the questionnaire, which seems to have been

sufficient, since there was no tendency that items in the latter part were unchecked.

11. In the final exam (5 hours written), there was one modelling task (30%) and one review task (20%) – both requiring skills that the peer-review exercises hopefully helped the students to develop.

Lecture attendance is not compulsory at the NTNU. There is no guarantee that those who did not attend lectures compensated for this by reading related compendium material. Hence, the students will have gone into the modeling and reviewing exercises with quite varying levels of preparation.

4.3 Data collected

The following data were collected from the activities mentioned in the enumerated list of section 4.2:

- The 20 case descriptions written by teaching staff.
- The 388 (2 x 194) conceptual models (one APM model per student, one Referent model per student).
- The 1164 (3 x 2 x 194) review reports (3 review reports per model). These consisted partly of numerical data (assigned scores for the model), and partly of free text data (identified defects and general comments to the model).
- The 66 responses to the questionnaire investigating the students' perceptions about their learning. As can be seen the response rate for the questionnaire was only 34% (not all students were present when the form was handed out, and not all present handed it in). The design of the questionnaire instrument will be discussed shortly.
- The final exam answers and grades. As with most Norwegian university exams, two independent censors do the grading, one internal (i.e., teacher) and the other external (i.e., not employed at the university). Then they meet do agree on the final grades. Unfortunately, it is impossible for us to correlate exercise performance (identified by student name) with exams (identified by anonymous numbers). The coupling between names and numbers is only known by the student administration, not to be revealed to teachers or scientific personnel.

We developed an own questionnaire to investigate the students' perceptions of their learning, because we did not find any existing instruments satisfactory for this task. There is a standard end-of-course questionnaire in use at the NTNU, but this suffers from many of the weaknesses that have been stated for such forms in general (Seymour, Weise et al. 2000; Snare 2000):

- "One size fits all", i.e., to be used in all courses taught at the university. As a result, the questions are not adapted (or adaptable) to the particular course being evaluated or teaching methods used.
- There is little connection between changes in teaching and the ensuing ratings, and teachers are often frustrated that standards course evaluation instruments measure how well the students liked the course rather than how well they learnt from it.

There are some published instruments that could have been more appropriate, e.g., SALG (Seymour, Weise et al. 2000) or "Student Opinion Survey of the Learning Process" (Snare 2000). But these had weaknesses in terms of survey design. In particular, the use of standalone survey items (observed variables) without underlying theoretical constructs (latent variables) makes evaluation of validity and reliability of empirical indicators difficult. While there have been a number of factor analytic studies used to identify underlying constructs used in course evaluation surveys (Feldman 1989; Abrami and d'Apollonia 1990; Marsh and Dunkin 1992), this represents *post hoc* definition of structure which is less desirable than *a priori* theoretical design of instruments (Pedhazur and Schmelkin 1991).

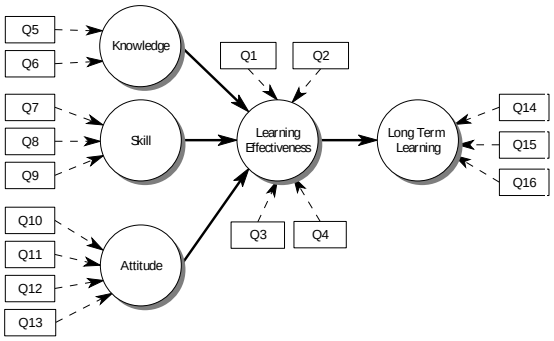

Figure 5. Underlying survey framework

The survey instrument, as shown in Figure 1, was meant to be generally applicable for evaluating the effectiveness of learning interventions (i.e., could also be customized to other courses than Information Systems and other learning interventions than peer reviews). More discussion about the general aspects of the instrument can be found in (Moody and Sindre 2003). The instrument is based on Bloom's taxonomy of educational objectives (Bloom 1984), i.e., that learning is composed of changes in knowledge, skill, and attitude. Moreover, it encompasses a distinction between short-term learning (i.e., for the exam) and long-term learning (i.e., for later courses and work life). The five circles in the diagram represent theoretical constructs (*latent variables*), while the arrows represent hypothesized causal relationships between them. Hence, the hypotheses are that positive contributions to Knowledge, Skill and Attitude will all be positive contributions to Learning Effectiveness (short-term learning), which will again be a positive contribution to Long Term Learning.

For each of the latent variables, we then developed observed variables (i.e., items for the questionnaire), indicated by the rectangles in Figure 1. In any customisation of the instrument, the observed variables must be based on the learning goals of the course and

intervention to be evaluated. The learning goals of SIF8035 were defined as follows:
Knowledge:

- K1: Understand the concepts of the modelling languages taught
- K2: Understand the concepts of the conceptual modelling quality framework

Skills:

- S1: Be able to use the modelling languages to develop conceptual models
- S2: Be able to interpret conceptual models
- S3: Be able to evaluate the quality of conceptual models

Attitudes:

- A1: Understand the importance of quality assurance in conceptual modelling
- A2: Understand the importance of conceptual modelling in the systems development process

These goals were used to develop items for the Knowledge, Skill and Attitude constructs. In addition, two general attitude items were defined, which evaluated the effect of the intervention on participants' enthusiasm/motivation for the course and their enjoyment of the course. Two additional items were defined for Learning Effectiveness, evaluating the effect of reviewing vs. being reviewed. All the items related to Figure 1 were to be answered in a 5 point Likert scale format.
In addition to the questionnaire items for the constructs of Figure 1, some questions related to the review process were also included in the questionnaire:

- One 5-option question asking how satisfied they were with the review feedback they received.
- Two questions investigating how much time was spent on review activities.
- 5 YES/NO questions about their preferences as to how the review process should be conducted.
- One open question asking for suggestions for improvement.

There were no hypothesized relationships between the process improvement part and the constructs shown in Figure 1. Since the process improvement part mostly investigates student opinion on what *should have* taken place, while the questions based on Figure 1 investigate what *did* take place, they are not easily correlated.
The questionnaire was written in Norwegian and presented to the students on four pages of A4 paper, to be answered by pen or pencil. An English translation of the questionnaire is shown in Appendix A.

4.4 Data Analysis Methods

In section 1 we posed four questions that might be interesting to ask. In the direct analysis of the data, we concentrate on the first two of these questions. The first question was: *Would it have been reasonable to use the scores from peer-review for grading?* This can be rephrased in several ways: Were the students reliable graders? Did good models receive good grades, and poor models receive poor grades? Are there signs of cheating, i.e., students copying models off other students? These questions will be investigated by:

- Statistical analysis of the Likert scale scores given during peer-review. Since each model got 3 independent reviews, we can simply look at the inter-rater reliability of these scores.
- Qualitative analysis: An expert looked through models and reviews, considering whether grades were reasonable, and whether there were any suspiciously similar models (possible cheating).

The second question was: *Did the introduction of peer-reviews improve the students' learning in SIF8035 Information Systems?* As mentioned earlier, this is hard to answer definitely, as we do not know what learning would have occurred if something else had been done instead of the peer reviews. Hence, the question must be rephrased into various sub-questions that can more easily be answered:

- Does the review reports indicate that the reviewers learnt something while making the reports? (e.g., do they contain clear applications of course knowledge?) And was the feedback of such an instructional quality that the modellers might in turn learn from it? Both these questions will be answered by a qualitative investigation of the models and reviews (the reviews are what really concern us here, but to say something about their quality, the models must also be looked at, together with the case descriptions they are based upon).
- Does their exam performance indicate learning. As observed earlier, we cannot correlate exam performance with exercise performance; neither does it make much sense to compare with exam results of the year before. Hence, only limited conclusions can be drawn from the exam.
- Do the students themselves think that they have learnt well from the peer review exercises? This will be answered by quantitative analysis of their responses to a questionnaire.

For qualitative analysis it would be too much work to look through 388 models and 1164 review reports. Hence we selected a random sample of 64 models (32 process models, 32 information models) and the 192 corresponding reviews as input for qualitative analysis. An expert then looked through this material. The expert holds a PhD in conceptual modelling, with 10 further years of research experience in the topic after graduating. The expert is also one of the inventors of the quality framework that the students applied for the reviews. Still, it must be admitted that the qualitative observations are unreliable. As long as only one expert is used, the degree of expert reliability is uncertain (Rust and Cooil 1994). Hence, it would have been better to use several experts, performing independent evaluations. However, the objective of the investigations was not to determine the exact quality of the students' work (for instance in terms of frequencies of various types of

errors), but to see if they might have learnt from it. Moreover, as will be seen in the following section, many of the expert-produced data are based on mere counts, not requiring expert judgment. This makes an accurate estimation of expert reliability less crucial.

5. RESULTS

5.1 Reliability of Scores

Each review report included 4 scores for various aspects of the model's quality, each in the range 1 (poor) to 7 (excellent). We could then look at the inter-rater reliability for the various scores of the same model, yielding the results shown in table 1. The reliabilities are around 0.6 for all the quality scores. (Nunally 1978) states that reliabilities should be at least 0.7, and specifically for grading there have been claims for reliabilities exceeding 0.8 or 0.9 (Walsh and Betz 2001; Kubiszyn and Borich 2003). Hence the scores from the 2002 peer reviews were not sufficiently reliable.

Table 5. Reliabilities for Each Quality Category

CONSTRUCT	CRONBACH'S α
Syntactic Quality	.6159
Semantic Quality	.5778
Pragmatic Quality	.5682
Overall Quality	.6091

5.2 Qualitative Analysis Of Review Scores

The impression that the scores were unreliable and would not have resulted in fair grades was also supported by the qualitative analysis. Some reviewers used much of the grading scale, for instance giving 2's and 3's to models they felt were poor, 5's – 7's to good models, in many cases matching the expert opinion about the models fairly well. Other reviewers used the scale quite defensively, typically giving 4's and 5's to all the models they reviewed, although some of these were significantly better than others. Hence some poor models undeservingly received mid-level scores, and some good models similarly. Some poor or mediocre models (in the expert's opinion) even received the highest scores (6's and 7's), the review comments indicating that the reviewers over-valued simplicity, favoring models that were far from complete (e.g., only modelling a minor part of what was described in the case). On the other hand, there were no observed cases in the sample where a good model received the lowest scores (e.g., 1's and 2's), so in cases where really low scores were given, the expert agreed that the model was poor.

5.3 Qualitative analysis of models

The quality of the models as such is not particularly important in assessing the learning outcomes of the peer-review exercises. But to assess the quality of the peer-review performance, it is necessary to have an idea of how many defects the models contained (that should have been found and reported by the students). The models were indeed varying a lot in quality. The best few had only 2-4 minor defects, while the worst easily had 10 or 20 defects, some of them quite serious. Typical defects for the process models were lacking flows/connections, missing roles, missing resources, lacking or wrongly specified decision points, poor naming of tasks, missing tasks, and wrong order of tasks. For the information models, the most frequent mistakes were cardinality errors on relationships, lacking attributes, lacking entity types, lacking relationship types, erroneous use of aggregation, membership and generalization connectors, lacking names for relationship types, system components modelled as entity types, e.g., "System" or "Database", and activities modelled as entity types. Hence, there were plenty of defects in the models that could have been discovered in the peer-reviews.

Another question during qualitative analysis of the models was the amount of copying. Here, the findings were quite encouraging: Only 4 (2 + 2) of the 64 delivered models in the inspected sample (e.g., about 6%) were so similar that the originators were unlikely to have worked independently. Whether the reason was copying or joint work is hard to say. Anyway, this finding indicates that the scheme with 20 different case descriptions (virtually 200) was successful in reducing the amount of copying. In previous years, when all students have used the same case description, copying has been a much bigger problem, maybe approaching 50%.

5.4 Qualitative Analysis Of The Review Defects Sections

In addition to the scores, the review reports contained a section where model defects should be identified. Typical quality figures for the identification of defects are:

- Ratio of Type 1 errors (false negatives or errors of omission): Defects exist in the model but are not identified by the reviewer. This ratio was about 60% for the process model reviews, 70% for the information model reviews.
- Ratio of Type II errors (false positives or errors of commission): Defects are claimed, but are not really defects. This ratio was about 5% for the process model reviews, 10% for the information model reviews.
- Ratio of Type III errors (classification errors): Defects are correctly identified but classified in the wrong category, according to the quality framework used (Lindland, Sindre et al. 1994). This ratio was 10% for the process model reviews, 20% for the information model reviews.

The above figures, however, are not necessarily useful, since they average out a wide range of student behaviours, from those doing virtually nothing, to those

putting a lot of effort into their reviews. Hence, it is interesting to look at various types of review performances. For the process model reviews (item 6 in the list of activities undertaken by the students, section 4.2) 21% reported 0 defects, 35% reported 1-2 defects, 31% reported 3-4 defects, and 13% reported 5 or more defects. For the information model reviews (item 9 in the list of activities undertaken by the students, section 4.2) 36% reported 0 defects, 27% reported 1-2 defects, 20% reported 3-4 defects, and 16% reported 5 or more defects.

Those reports that identified most defects were also the most likely to demonstrate significant understanding of modelling principles. For those who reported zero defects, it is impossible to see any evidence of learning from the review reports, especially since these reports also tend to give defensive scores. A coarse observation could be that those who delivered fairly thorough review reports revealing application of modelling knowledge are likely to have learnt from performing the reviews, the other students not.

The varying quality of work by the reviewers of course had direct impact on what the reviewees received in terms of feedback. Assuming that it would not be too bad for the reviewee if two of the reports were meagre (e.g., reporting zero or few defects) as long as the third was thorough, we group the models according to the review identifying most defects. For the process models,

- 6% of the students received no reported defects (all 3 reports blank in the defects section).
- 19% had 1 or 2 defects in the most thorough report.
- 37% had 3 or 4 defects in the most thorough report.
- 37% had 5+ defects in the most thorough report.

And for the information models,

- 16% had no reported defects.
- 22% had 1 or 2 defects in the most thorough report.
- 28% had 3 or 4 defects in the most thorough report.
- 34% had 5+ defects in the most thorough report.

As students were not required to correct the defects and resubmit the model afterwards, it is hard to establish how much the reviewees actually learnt from the feedback. But at least, those who received thorough reviews had the *opportunity* to learn from it. Those who received few or no hints of their defects, had less opportunity.

5.5 The Exam Performance

At the NTNU, two persons determine the grades. They look through the answers independently, scoring them 0-100, then meet to compare scores and decide on grades (A-F). The exam assignments most closely related to the exercises were #1 (making an APM model, 30 points) and #2 (reviewing a use case description, 20 points). For #1 the average score was 75%, for #2 it was 69%. #2 contained a use case description with some deliberately introduced defects, the students' task being to identify the defects and evaluate the overall quality of the use case description. Most of the students were able to identify at least 50% of the defects. This is a lot better than their exercise performance, where many identified few or no defects. Also, their average modelling score of 75% seems a lot better than their modelling performance in the exercises, when many models were just good enough to pass (around 40%).

5.6 Reponses to the Questionnaire

Analysis of the instrument itself: The instrument was investigated for construct validity and reliability. All items were found valid. As shown in Table 6, two of the constructs (Learning Effectiveness and Long Term Learning) had high levels of reliability (> .7), while the constructs associated with the learning goals had lower than acceptable levels. This indicates that more care needs to be taken in formulating learning goals to ensure that they are clear and precisely defined.

Table 6. Item Reliabilities

CONSTRUCT	CRONBACH'S α
Knowledge	.432
Skill	.640
Attitude	.642
Learning Effectiveness	**.773**
Long Term Learning	**.855**

Evaluation of Latent Variables: Table 7 shows the summary statistics for each construct. Overall, students found the review exercises to be moderately effective in improving their knowledge, skills and attitude and their learning in the course, but only between slightly and moderately effective for long term learning. While these results are encouraging, there is clearly room for improvement—this represents a "lukewarm" response rather than an overwhelmingly enthusiastic one. If we take 3 as the break-even point, the only item which is significantly greater than 3 is Skill ($\alpha < .05$).

Table 7. Summary of Construct Values

CONSTRUCT	MEAN	STDEV	RESULT
Skill	3.28	.68	Moderate
Knowledge	2.93	.72	Moderate
Attitude	3.01	.63	Moderate
Short Term L.	3.13	.61	Moderate
Long Term L.	2.54	.85	Slight

Observed Variables: Table 8 summarizes the responses to the individual survey items in descending average

scores (i.e., from most positive to least positive). It must be noted that for the three questions Q4, Q10, Q11, the "zero" score is 3, as 1 and 2 constitute negative options, for instance that the peer review exercises reduced the enjoyment of the course (Q11). For the other questions, score 1 is the "zero" score, e.g., to state that nothing was learnt from the review exercises (Q1). Some conclusions to be drawn:

- The students definitely seem to think that they have learnt from the peer review exercises (if they had not, their answers should have been much closer to 1). However, since averages are mostly in the middle range, the response is not overwhelmingly positive.
- A crucial question with regards to whether peer reviews have improved the course may be Q4, asking whether the learning from the peer reviewing was obtained more (or less) effectively/efficiently (in Norwegian, the word "effektiv" as used in the questionnaire means both effective and efficient) than it would have been from other suggested techniques (e.g., lectures, self-study, more modelling without peer-review). The response is above 3, but only slightly so, which makes it difficult to draw strong conclusions. Since this question may be particularly interesting, we look more in depth at how the answers were distributed along the scale:
  - 1 student: peer reviews were far less effective.
  - 17: slightly less effective
  - 14: as effective as other learning techniques
  - 27: slightly more effective
  - 7: much more effective

  Hence, there were 34 who thought that peer reviews were pedagogically more effective than the alternatives, while only about half that amount (18) thought they were less effective, and only 1 strongly so. While this does not prove anything, it at least suggests that the students think peer reviews may have been an improvement, and that they would have learnt less from the course if the review activities had not been introduced.
- Students did not see a great benefit from the review exercises beyond the course itself, as the scores for long-term learning are fairly low. This suggests that more effort could have been spent explaining to students the relevance of reviews in future work.
- Students found the process of reviewing others' models more useful than the process of being reviewed. This is perhaps not surprising. Since it was not compulsory to correct the models after the review, most students did not do this, thus not learning much from any feedback.

Determinants of Learning: A number of causal relationships were hypothesized between the constructs in the theoretical model:

- Knowledge + Skill + Attitude → Short Term Learning
- Short Term Learning → Long Term Learning

Table 8. Summary of Item Responses

Item	Constr.	Mean	St D
Q10: Enthusiasm for course	*Attitude*	*3.36*	*.89*
Q6: Understanding of quality framework	Knowledge	3.33	.85
Q4: Relative ped. effectiveness of review	*Short Term Learning*	*3.33*	*1.04*
Q9: Ability to evaluate quality of models	Skill	3.30	.82
Q11: Enjoyment of the course	*Attitude*	*3.32*	*.71*
Q5: Understanding of modeling languages	Knowledge	3.23	.86
Q12: Importance of QA in conceptual modeling	Attitude	3.17	1.03
Q2: Learning from reviewing others'	Short Term Learning	2.91	.89
Q8: Ability to interpret conceptual models	Skill	2.86	.88
Q7: Ability to develop quality models	Skill	2.85	.77
Q1: Overall learning from review exercises	Short Term Learning	2.85	.89
Q13: Importance of modeling in IS dev.	Attitude	2.67	.91
Q15: Preparation for project work	Long Term Learning	2.65	.98
Q3: Learning from reviews by others	Short Term Learning	2.64	.94
Q16: Preparation for working life	Long Term Learning	2.56	.96
Q14: Preparation for future courses	Long Term Learning	2.39	.96

These causal relationships were confirmed by multiple regression analysis. The first with $\alpha < .01$, the adjusted r^2 statistic showing that together the learning goals accounted for 44% of the variance in Short Term

Learning. More detailed investigations of the findings indicated that Skill and Attitude had a significant effect on Long Term Learning while Knowledge did not. These details are not particularly important to the research questions of this publication, but are discussed in (Moody and Sindre 2003).

Table 9. Responses to Process Questions

QUESTION	Y%	SIGN	
As a Reviewer			
Q17: Anonymity of reviewee	5%	.000**	**Y**
Q18: Corresp. w/ reviewee	58%	(.268)	?
Q19: Coop w/ other reviewers	67%	.010*	**Y**
As a Reviewee			
Q20: Anonymity of reviewer	20%	.000**	**Y**
Q22: Respond to reviewer	74%	.000**	**Y**

Time spent: The questionnaire also included two questions on the time spent for the exercises, one about the total time spent on performing peer reviews, i.e. activities 6 + 8 in the list in section 4.2, and one about the time spent looking at the feedback and (possibly) improving ones own model based on this. No hypothesis was made between this and the constructs in the theoretical model, but knowledge of the time spent could still be useful in interpreting the results. On average, the students who answered the questionnaire spent 3.6 hours performing reviews, which is less than 50% of the time nominally allocated (4 hours for the process model reviews + 4 hours for the information model reviews). Only 1 of the 66 students answering the questionnaire reported spending more than the allocated 8 hours. 17 of the 66 students (about 25%) spent 2 hours or less, which gives less than 20 minutes for each of the 6 reviews performed. Within such a 20 minute time frame, the student would have to read and understand the case description, then look at the model and determine how well it corresponded to the case description, identify and report defects and assign scores. It is thus no surprise that many reports failed to identify defects or only picked out one or two obvious ones. Furthermore, only 1.2 hours were spent looking at feedback to own models (i.e., approximately half an hour for each of the models). But there was no compulsory activity requiring further work based on the feedback, and much of the feedback did not contain significant information anyway. Hence a limited spending of time on this latter activity is easily understandable.

Process Improvement: For the 5 Yes/No questions the Binomial Test was used to determine whether the responses to each question were significantly positive or negative. Table 5 summarizes the results of the significance testing. Overall, participants wanted reviews to be anonymous (as they were), but would have liked to collaborate with other reviewers and have the ability to respond to reviewer comments (which was not possible this time around). Qualitative responses to the open question Q25 were also analysed. Many students did not make any personal suggestions for improvement, hence the percentages for various types of suggestions are fairly low. The most common suggestions were: "Improve the web-based evaluation system" (12%), "Requirements to pass the review exercises should be stricter" (9%), "Should have reviewed the same case as we modeled" (8%), "Should have had more iterations" (4%), i.e. modeling, getting reviews, improving the model based on reviews, getting reviewed again, ... "Lectures should have been more relevant" (4%). The fact that many students left this question blank cannot be taken to mean that they were totally satisfied with the review process (i.e., saw nothing to improve), as lack of an answer may also be due to lack of time or motivation for answering the question, or lack of concrete ideas for improvement.

6. DISCUSSION

In this section we will discuss each of the four questions that were posed in the Introduction, then finally discussing some threats to the validity of our conclusions.

6.1 Scores Usable For Peer Grading?

Our first research question was: *Would it have been reasonable to use the scores from peer-review for grading?* The answer to this is no (contrary to our hopes). The inter-rater reliability was as low as 0.6, which is not considered acceptable. True, in a normal two-censor situation, the resulting grades may be fair in spite of low reliabilities, for instance if one censor is systematically too strict and the other systematically too generous, which then averages out. In a peer-grading situation, however, there are as many graders as there are students, so generally, discrepancies will not be averaged out. It can easily happen that some students get 3 generous reviewers and others get 3 strict reviewers. Moreover, some students did not use the scale very much, defensively assigning 4's and 5's to all the models they reviewed, although these models varied significantly in quality.

Why did the students grade so inadequately? There are several explanations for this:

- Too little time spent, as indicated both by qualitative investigations and the questionnaire responses.
- Lack of motivation. There were no rewards for fair scoring and no punishments for inaccuracy.
- Unclear instructions from staff. The students did not receive detailed instruction on the meaning of

the various scores or what should be deducted for various types of model defects. Neither were there clear standards for the expected detail levels of the models delivered. Hence, it was difficult for the students to grade reliably, even if they had spent more time than they did.
- Limited skills. It was the students' first tries at making APM models and Referent models, and similarly their first attempts at reviewing such models.

6.2 Learning Improvement?

Our second research question was: *Did the introduction of peer-reviews improve the students' learning in SIF8035 Information Systems?* This question cannot be definitely answered based on the available material. The following observations might suggest a NO:

- A significant amount of the delivered review reports contained little or no evidence of any learning (qualitative investigations of review reports).
- On average the students spent much less time performing the reviews than they were supposed to (answers to questionnaire).
- A significant amount of the students received feedback that they could not possibly learn from (qualitative investigations of review reports).
- Only limited time was spent looking at review feedback (answers to questionnaire).

However, other observations suggest a YES:

- About half of the students delivered review reports that successfully applied curriculum knowledge to identify defects (although not finding all defects). Even the knowledge applied might have been gained before the review activity took place, the application of knowledge in a review situation is useful training in its own right, thus a sign that learning has taken place.
- Also, about half of the students received review reports containing useful insights on model defects. But few took the time to improve their models, so less learning can be claimed from this.
- The students performed reasonably well with the modelling and review questions in the exam. So they must have learnt something during the term.
- The responses to the questionnaire indicate that the students feel that they learnt from the peer review exercises. Q4 shows 34 of 66 respondents thinking that peer reviewing was more effective than other learning activities for some of the course learning goals. Only 18 of 66 felt it was less effective.

All in all, this points to the conclusion that the peer review activities did improve the learning in the course for some students, but far from all. Typically, students who bothered to make detailed review reports are likely to have learnt from it. So are those who were lucky to receive detailed review reports on their own models and took the time to revisit their model to understand the reviewer comments. On the other hand, students who spent minimal time on the activities, only doing enough to pass, did not learn much. For these students, the review exercises have not led to improved learning – except maybe if they revisited the exercises during the final weeks before the exam. This is quite common to do at the NTNU, since exercise topics may indicate what the teacher considers important in a course.

6.3 Ideas for Improvement?

Our third question from the introduction was: *Would it have been better if peer-reviews had been introduced in another way?* Our findings indicate some positive and some negative outcomes of the 2002 review exercises. The major challenge seems to be to motivate the students better, so that all (or at least nearly all) make honest attempts at the task. Some possible carrots:

- Make a stronger case for reviews in later work life (e.g., through lectures and the selection of reading materials), possibly making the exercise review process more similar to industrial review processes (especially if a review-as-topic profile is wanted). In 2002, there were 20 pages about validation of requirements specifications (Kotonya and Sommerville 1997), including discussion of requirement reviews. However, this excerpt only treats the subject quite superficially, and was lectured after the review exercises had taken place.
- Use modelling languages that are common in industry, to increase the relevance for work-life.
- Emphasize more strongly that review questions are likely for the exam.
- Introduce reviews more gradually, helping the students reach some level of mastery (e.g., plenary classroom exercises reviewing smaller models with some deliberately introduced defects?) before they are given bigger models with no definite solution.
- Enable for more group discussion about the reviews (e.g., the 3 reviewers of the same models meet afterwards to arrive at a consensus decision), resembling review meetings in industry.
- Make the task easier for the students. For instance, let the three models that a student is supposed to review be according to the same case description. Then the student only has to relate to one new case description during review rather than three, and there will be something to compare with respect to defects and scoring.

And some possible sticks:

- Make more detailed requirements on expected review output, and flunk those who do not comply.
- Force the students to correct their models and resubmit them after the reviews have been received.
- Force the students to perform the reviews in a lab at fixed times, supervised by teaching staff.

Not all these measures are equally tempting (for instance, if the review process is to be supervised by

staff, this would increase the workload a lot). Also, improvement must be seen in relation to a certain purpose. With review-as-topic, the emphasis on work-life relevance would be the most important. With review-as-pedagogy, it might be more interesting to stress team-work and iterative improvement (e.g., that reviews are discussed in a team, and that the modeller has to improve the model and resubmit for new reviews). For peer-grading, the major challenge will be to make the scores more reliable. Means to increase the reliability might be to:

- Provide more instruction on grading, i.e., what qualities (or lacks thereof) would be expected in models receiving various scores, grading some example models in the classroom, etc.
- Train the students' grading capabilities before it gets serious, e.g., trial grading assignments assessed by staff.
- Ensure a certain level of curriculum knowledge relevant for grading. In 2002 students had one shot at making an APM model and a Referent model. Using modelling languages known from before, or giving them several exercises with the same language, their modelling and reviewing capabilities might have been better.
- Have some quality assurance mechanisms, such as teaching staff looking through grades in cases of unacceptable discrepancies between various student reviewers.
- Provide motivational mechanisms, e.g., possible grade punishments for sloppy / inaccurate grading.

One major weakness of the 2002 approach may have been that the purpose of peer reviews was not fully clarified beforehand. For instance, our approach had some aspects of review-as-topic and some of review-as-work-reduction, but failed to comply fully with any of them (e.g., too little focus on industrial review processes for the topic angle, too little facilitation of reliable grading for the work-reduction angle).

6.4 Guidelines for Peer Reviews in IS Courses

Our fourth and final question in the Introduction was: *Should peer-reviews be included in IS Analysis and Design courses in general, and what teaching guidelines can be extracted from our experience?* We feel that courses with a how-to profile (i.e., how to develop information systems, dealing with issues such as modeling techniques, problem statements, and requirements specifications) should have some focus on peer-review, if nothing else for its relevance as a topic in its own right. As discussed previously, it can also be motivated pedagogically or for reducing the workload of teachers. The main guidelines from our experience are the following:

- Decide beforehand the main purpose of the peer reviews (e.g., topic, pedagogy, peer-grading). It is legitimate to have several motivations, but if so the various trade-offs must be worked out to ensure that at least the main purpose is fulfilled.
- Whatever the purpose, a major challenge is to motivate the students. Otherwise, they may produce unreliable grades and useless feedback, and not learn what they were supposed to.
- Provide clear requirements as to what is expected from the reviewers. In 2002 the requirements were unclear, and students easily passed without identifying defects.
- There should also be clear quality standards for the artifacts that are subject to peer-review (in our case: the conceptual models). Not only will this help the students as reviewers, it will also help them producing artifacts of a better quality *before* review.

As concluded by another study, the quality framework that was used during review in 2002 is quite abstract, not customized to the modeling languages used, and thus of limited help to the students in identifying defects (Moody, Sindre et al. 2002; Moody, Sindre et al. 2003). For future offerings this might be supplemented with quality standards particularly designed for the type of model to be made.

6.5 Threats to Validity

As mentioned before, only one expert was used for the qualitative investigations, and this expert may be unreliable. One may therefore contend that the use of several independent expert evaluations would have been better, which we also agree. This was not done because the qualitative investigations took a lot of time, so few were tempted to undertake them. But in our opinion, possible errors in expert judgment are not likely to be a problem with regards to the conclusions to be drawn. First of all, many of the data reported by the expert were obtained by mere counting (e.g., how many reports mentioned 0, 1, 2, ... defects), not demanding any judgment of right/wrong or good/bad. Moreover, for most defects that *were* reported by the students, there was agreement between the expert and the students (cf. the low ratio of Type II errors). Hence, what the expert mostly criticized in the students' work was failure to report defects and arbitrary grading. Both these are supported by observations not made by the qualitative expert (the low time spent, and the low reliability of grades).

Another weakness is the low response rate to the end-of-course questionnaire (34%), which may have caused a bias. For instance, those students who attended the last lecture and handed in the questionnaire might be more positive to the course in general (and thus also to the peer review exercises) than those who did not attend, or attended without handing in the questionnaire. This may have caused too positive conclusions about the learning effectiveness of peer reviews. But many of the respondents to the questionnaire had spent little time on the peer review exercises. So the responding sample was certainly not uniformly composed of students who had been strongly motivated and worked hard with the

exercises.

Another issue is that the questionnaire might have given better data if there had been more items or the questions had been more precise. Moreover, some questions possibly had a too high granularity. For instance, the questions that asked the students about the time spent did not separate between the two review exercises but asked the students to report the total spending. Similarly, Q4 asks about the relative learning effectiveness of peer reviews versus various other learning techniques (lectures, self-study of compendium, modelling). With separate items for each of these the picture could have been more complete. However, the inclusion of more questions would have made the questionnaire longer, and this is a difficult balance.

A final concern may be the generalizability of our conclusions. The observations have been made for one offering of one course, in one particular university, mostly with Norwegian students in the age group 21-25. Hence, in another kind of course, with another learning culture and different students (for instance well motivated in the first place), the outcome might have been different, and some of our hindsight guidelines less suitable.

7. CONCLUSION

This paper has analysed the introduction of peer review in the course SIF8035 Information Systems at the NTNU. The course has approximately 200 students per offering and its main focus is conceptual modelling in IS analysis. Peer reviews were performed in connection with two compulsory modelling exercises in the course. In both these, each delivered model was double-blind reviewed by three independent peers. The review reports contained quality scores for the model and a list of identified defects explained in free text.

The main question of interest was whether the introduction of peer reviews improved the course. As the peer reviews were part of the mainstream teaching, the teaching was not done in the form of a controlled experiment, and there is no "control group" who did not receive the peer review treatment. Hence it is impossible to use with/without comparisons to confirm improvement according to peer reviews. Moreover, as argued earlier in the paper, such comparisons are very hard to make, and in some cases unethical or against university regulations. Instead we choose to assess the learning effectiveness of peer reviews based only on investigating the performance and perceptions of the students who took part in this offering of the course. The complexity of the research question then called for the triangulation of various methods:

- Quantitative analysis of student performance, looking at the inter-rater reliability of peer scores.
- Qualitative analysis of student performance, as an expert looked through sample models and reviews.
- Quantitative analysis of student perceptions, from an end-of-course questionnaire.
- Qualitative analysis of student perceptions, from an open question in the same questionnaire.

In spite of some threats to validity (e.g., use of only one expert for qualitative investigations, limited response rate for questionnaire, no control group), we have argued that the following conclusions can be drawn:

- The scores assigned in the 2002 peer reviews were not reliable enough for grading.
- Some performed the review tasks reasonably well, and may have learnt much from them. Others delivered thin reports, showing little sign of learning.
- Similarly, some got useful feedback from the peer reviews, which could help them improve their modelling skills. Others received feedback that would not have been useful in this respect.
- The students felt that they have learnt from the peer review exercises, but the reception is only lukewarm. A slight majority of the students felt that peer reviews were more effective than other available teaching methods in achieving mentioned learning goals, but only marginally so, not mandating strong conclusions.

To sum up, the introduction of peer reviews may have improved the students' learning in the course, but not equally much for all students. Those who did little, learnt little, causing some quite disappointing figures in the analysis of the review performances. Still, we feel that experiences are more positive than negative, and it is planned to continue with peer reviews for the next offerings of the course. Based on hindsight, we have provided a discussion of the trade-offs between three different purposes of peer-reviews – this may serve as a guideline to us (and hopefully others) in the future.

8. FURTHER RESEARCH

The most obvious scene for further research on this topic (at least by these authors) would be later offerings of the same course. Then, it would be natural both to improve the course itself and the research methods. Concerning improvements to the course, the natural first step would be to follow our own guidelines: Decide on the main goal of the peer reviews and do what is necessary to facilitate it. Anyway, measures must be taken to motivate the students better. This might suggest a stronger focus on review-as-topic or review-as-pedagogy, rather than review-as-work-reduction (the latter less popular with students, not wanting burdens they feel should have been the teacher's?). But given the current situation at Norwegian universities, with a huge number of students per teacher (especially in informatics departments), the teaching staff cannot evaluate hundreds of conceptual models and review reports in a thorough manner. Yet a complicating issue is the *Reform of the quality of higher education* (UFD 2001),

that was passed by the Norwegian parliament (*Storting*) in 2001. According to this reform, feedback to students should be more frequent and course grades partly based on work during the term, not on a final exam alone. This may force the peer-grading angle. A major challenge, then, will be to make the students grade adequately.

Concerning the research method, the results could be stronger if several experts were used for the qualitative analyses, so that their reliability could be estimated (Rust and Cooil 1994). One could also take measures to secure a higher response rate to the questionnaire. An additional advantage of subsequent research is the possibility of comparing questionnaire answers from one year to the next. This might discourage changes to the questionnaire, but some improvements should be considered. The questions based on defined learning goals had too low reliabilities in the 2002 investigation.

The questions were based on the knowledge/skill/attitude taxonomy of (Bloom, 1984), but disregarded the more detailed levelling within these three categories. For instance, within knowledge (cognitive domain) Bloom's taxonomy has the following levels:

- **Knowledge:** repeating from memory
- **Comprehension** of terms and concepts
- **Application** of information to solve a problem
- **Analysis:** breaking things down into their elements, formulating theoretical explanations or models for observed phenomena
- **Synthesis:** creating something, combining elements in novel ways
- **Evaluation**: choosing from among alternatives and justifying the choice using specified criteria

Such levels, or similar inspirations from other taxonomies, e.g., (Gagne, Briggs et al. 1992) could hopefully inspire a more precise definition of learning goals and questionnaire items. Notably, peer-review itself stimulates the achievement of advanced learning goals like "evaluation" on Bloom's scale. This might in itself serve as a weighty argument that this teaching method has a place in IS Analysis and Design courses like SIF8035.

9. REFERENCES

Abrami, P. C. and d'Apollonia [1990]. The Dimensionality of Ratings and Their Use in Personnel Decisions. Student Ratings of Instruction: Issues for Improving Practice: New Directions for Teaching and Learning. J. Franklin. San Francisco, Jossey-Bass.

Andersen, R., R. Conradi, J. Krogstie, G. Sindre and A. Sølvberg [1994]. "Project Courses at the NTH: 20 years of experience." 7th SEI Conference on Software Engineering Education (CSEE'94), San Antonio, TX, 5-7 Jan. Springer Verlag.

Bancroft, N. H., H. Seip and A. Sprengel [1997]. Implementing SAP R/3: How to Introduce a Large System into a Large Organization. Greenwich, CT, Manning / Prentice Hall.

Bloom, B. [1984]. Taxonomy of Educational Objectives. New York, Longman.

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, May 27-31. 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, Aug 20-22. IEEE Computer Society Press.

Collofello, J. S. [1987]. "Teaching Technical Reviews in a One-Semester Software Engineering Course." ACM SIGCSE Bulletin **19**(1): 222-227.

Computas [2001]. What is METIS? http://www.metis.no/products/what_is_metis.html checked: 6-Aug, 2002.

Curran, T. A. and A. Ladd [2000]. SAP R/3 Business Blueprint: Understanding Enterprise Supply Chain Management. Upper Saddle River, NJ, Prentice Hall PTR.

Davis, A. M. [1995]. "Object-Oriented Requirements to Object-Oriented Design: An Easy Transition?" Journal of Systems and Software **30**: 151-159.

Eschenbach, E. A. [2001]. "Improving technical writing via web-based peer-review of final reports." 31st ASEE/IEEE Frontiers in Education Conference, Reno, NV, 10-13 Oct. IEEE.

Farschchian, B. A. [2001]. A framework for supporting shared interaction in distributed product development projects. Department of computer and information science. Trondheim, Norwegian University of Science and Technology.

Felder, R. M. [1992]. "What Do They Know Anyway?" Chemical Engineering Education **26**(3): 134-135.

Feldman, K. A. [1989]. "The Association Between Student Ratings of Specific Instructional Dimensions and Student Achievement: Refining and Extending the Synthesis of Data from Multisection Validity Studies." Research in Higher Education **30**: 583-645.

Fowler, M. and K. Scott [2000]. UML Distilled: A brief guide to the standard object modeling language. Reading, MA, Addison-Wesley Professional.

Gagne, R. M., L. J. Briggs and W. W. Wager [1992]. Principles of instructional design. Fort Worth, Harcourt Brace Jovanovich.

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, 21-25 Feb. ACM Press.

Gilb, T. and D. Graham [1993]. Software Inspection. Reading, MA, Addison-Wesley.

Gulla, J. A. and T. Brasethvik [2000]. "On the Challenges of Business Modelling in Large-scale Reengineering Projects." 4th International Conference on Requirements Engineering (ICRE 2000), Schaumburg, IL, 19-23 June 2000. IEEE Computer Society Press.

Hirschheim, R. and H. K. Klein [1989]. "Four Paradigms of Information Systems Development." Communications of the ACM **32**(10): 1199-1216.

Jick, T. D. [1979]. "Mixing Qualitative and Quantitative Methods: Triangulation in Action." Administrative Science Quarterly **24**: 602-611.

Johansson, C. [1997]. "Early Practise and Integration - The Key to Teaching Difficult Subjects." International Symposium on Software Engineering in Universities (ISSEU'97), Rovaniemi, Finland, 6-8 Mar.

Kelly, J., J. Sherif and J. Hops [1992]. "An Analysis of Defect Densities Found During Software Inspections." Journal of Systems and Software **17**(2): 111-117.

Kirchmer, M. [1999]. "Improve Business Processes Based on ERP and Post-ERP Applications." ERP World '99 Conference & Expo, San Francisco, August 17-19.

Kotonya, G. and I. Sommerville [1997]. Requirements Engineering: Processes and Techniques. Chichester, John Wiley & Sons.

Kubiszyn, T. and G. Borich [2003]. Educational Testing and Measurement. New York, Wiley.

Kulak, D. and E. Guiney [2000]. Use Cases: Requirements in Context. Reading, MA, Addison-Wesley.

Lindland, O. I., G. Sindre and A. Sølvberg [1994]. "Understanding Quality In Conceptual Modelling." IEEE Software **11**(2): 42-49.

Maiden, N. and C. Ncube [1998]. "Aquiring COTS Software Selection Requirements." IEEE Software **15**(2): 46-56.

Marsh, H. W. and M. Dunkin [1992]. Students' Evaluations of University Teaching: A Multidimensional Perspective. Higher Education: Handbook of Theory and Research Vol. 8. J. C. Smart. New York, Agathon.

Moody, D. L. and G. Sindre [2003]. "The Learning Effectiveness Survey (LES): An instrument for evaluating and improving the effectiveness of learning interventions." Hawaii International Conference on Education, Waikiki, 7-10 Jan.

Moody, D. L., G. Sindre, T. Brasethvik and A. Sølvberg [2002]. "Evaluating the Quality of Process Models: Empirical Analysis of a Quality Framework." 21st International Conference on Conceptual Modeling (ER'2002), Tampere, Finland, October 7-11.

Moody, D. L., G. Sindre, T. Brasethvik and A. Sølvberg [2003]. "Evaluating the Quality of Information Models: Empirical Testing of a Conceptual Model Quality Framework." 25th International Conference on Software Engineering, Portland, OR, 3-10 May.

Newell, J. A. [1998]. "The Use of Peer-Review in the Undergraduate Laboratory." Chemical Engineering Education **32**(3): 194-196.

Nunally, J. [1978]. Psychometric Theory (2nd edition). New York, McGraw Hill.

Parsons, J. and Y. Wand [1997]. "Using Objects for System Analysis." Communications of the ACM **40**(12): 104-110.

Pedhazur, E. J. and L. P. Schmelkin [1991]. Measurement, Design and Analysis: An Integrated Approach. Hillsdale, NJ, Lawrence Erlbaum Associates Inc.

Preece, J. J., Y. Rogers, H. Sharp, D. Benyon, S. Holland and T. Carey [1999]. Human-Computer Interaction, Addison-Wesley.

Rust, R. T. and B. Cooil [1994]. "Reliability Measures for Qualitative Data: Theory and Implications." Journal of Marketing Research **31**(February): 1-14.

Sawyer, P. [2000]. "Packaged software: Challenges for RE." Sixth International Workshop on Requirements Engineering: Foundation for Software Quality, Stockholm, Sweden, Essener Informatik Beiträge, Essen, Germany.

Scheer, A.-W. [1998]. Business Process Engineering: Reference Models for Industrial Enterprises. Berlin, Springer Verlag.

Scheer, A.-W. [1999]. ARIS - Business Process Frameworks. Berlin, Springer Verlag.

Seymour, E., D. J. Weise, A.-B. Hunter and D. S.M. [2000]. "Using Real-World Questions to Promote Active Learning." Proceedings of the National Meeting of the American Chemical Society Symposium, San Francisco, March 27.

Snare, C. E. [2000]. "An Alternative End-of-Semester Questionnaire." Political Science Online(December).

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. [2000]. Introduction to Concept Modeling for Information Systems. Norwegian University of Science and Technology, Trondheim, Norway.

Sølvberg, A. and D. C. Kung [1993]. Information Systems Engineering, Springer Verlag.

Sorge, M. [2000]. Evaluering av prosjektundervisningen ved Institutt for datateknikk og informasjonsvitenskap, NTNU. Technical report Program for lærerutdanning, Seksjon for universitetspedagogikk, NTNU.

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, 10-13 Oct. IEEE.

UFD [2001]. Do your duty - demand your right.

Norwegian Ministry of Education, Research and Church Affairs.

van der Aalst, W. M. P. [1999]. "Formalization and Verification of Event-driven Process Chains." Information and Software Technology **41**(10): 639-650.

van Vliet, H. [2000]. Software Engineering: Principles and Practice. Chichester, John Wiley & Sons.

Walsh, W. B. and N. E. Betz [2001]. Tests and Assessment. Englewood Cliffs, NJ, Prentice-Hall.

Wiegers, K. E. [2001]. Peer-Reviews in Software: A Practical Guide. Boston, Addison-Wesley.

AUTHOR BIOGRAPHIES

Guttorm Sindre in an Associate Professor in the Dept of Computer and Information Science, Norwegian Univ. of Science and Technology (NTNU). He received a PhD in information systems from the same university in 1990. His research interests include requirements engineering, conceptual modelling and IT education. Part of his work with this paper took place while visiting the Dept of Management Science and Information Systems, Univ. of Auckland, New Zealand.

Daniel Moody is a Visiting Professor in the Dept of Software Engineering, Charles Univ. in Prague (visiting from the School of Business Systems, Monash Univ., Australia). He has a PhD in Information Systems from the Univ. of Melbourne and has held positions at several universities in Australia and Europe, including the NTNU. He is the current Australian President of the Data Management Association (DAMA) and Australian World-Wide Representative for the Information Resource Management Association (IRMA).

Terje Brasethvik received his Master's degree in Computer Science from the NTNU in 1996. He was then employed in the oil company Statoil as a researcher, before he went back to the university to work on his PhD in 1997. His research interests are conceptual modelling, information systems engineering and CSCW.

Arne Sølvberg is professor of Computer Science at NTNU (NTH), Norway, since 1974. His main fields of competence are information systems design methodology, database design, and information modelling, advising 20 PhD students over the last 10 years. He has participated in several EU-sponsored research projects and been active in international organizations for research cooperation such as IFIP and the VLDB Endowment, as well as co-founding the CAiSE conference series. He has been a Visiting Scientist with IBM San Jose Research Labs, The Univ. of Florida, the Naval Postgraduate School in Monterey, and the University of California at Santa Barbara.

APPENDIX A. QUESTIONNAIRE

Note: For space reasons (fitting questionnaire on two pages instead of four), font sizes have been reduced, the format changed from one-column to two-column, and some answer options abbreviated where, of course, words were written in full in the original Norwegian questionnaire. Otherwise, the below is a straightforward translation of the Norwegian questionnaire that was handed out to the students. With respect to Q4 a slight translation problem should be noted: the Norwegian word "effektivt" (as used in the original's answer options) covers both the words "effective" and "efficient" in English.

LEARNING EFFECTS OF PEER REVIEW EXERCISES

Peer reviews were introduced in SIF8035 this year, so we wish to evaluate this teaching technique to find out whether it gave a useful learning experience. We are also interested in suggestions for improvement that can benefit next year's students. THERE ARE NO RIGHT OR WRONG ANSWERS TO THE QUESTIONS BELOW, AND YOUR ANSWERS WILL HAVE NO CONSEQUENCES FOR THE GRADING IN THE COURSE – PLEASE JUST GIVE YOUR HONEST OPINIONS.

PART 1. OVERALL LEARNING

Q1. How much did the review exercises contribute to your learning in this course?

□ Nothing □ A little □ Medium □ Much □ Very much

Q2. Did you learn something from reviewing others' models?

□ Nothing □ A little □ Medium □ Much □ Very much

Q3. Did you learn something from the feedback on your own models?

□ Nothing □ A little □ Medium □ Much □ Very much

Q4. How effective was the learning from the review activities compared to learning the same through more lectures, reading, modelling without peer feedback, or the like? "Learning through peer review was..."

□ Clearly less effect □ Less effect □ Equally effective □ More effect □ Clearly more effective

PART II: KNOWLEDGE

Q5. How much did the review activities contribute to your understanding of the modelling languages used?

□ Nothing □ A little □ Medium □ Much □ Very much

Q6. How much did the review activities contribute to your understanding of the concepts of the quality framework?

□ Nothing □ A little □ Medium □ Much □ Very much

PART III: SKILLS

Q7. How much did the review activities contribute to your ability to make process- and information models of high quality?

□ Nothing □ A little □ Medium □ Much □ Very much

Q8. How much did the review activities contribute to your ability to understand process- and information models made by others?

□ Nothing □ A little □ Medium □ Much □ Very much

Q9. How much did the review activities contribute to your ability to assess the quality of process- and information models?

□ Nothing □ A little □ Medium □ Much □ Very much

PART IV: ATTITUDE

Q10. How did the review activities affect your motivation for the course?

□ Strongly demotiv. □ Slightly demotiv. □ No effect □ Slightly motiv. □ Very motivating

Q11. What effect did the review activities have on your enjoyment of the course?

□ Strongly negative □ Slightly negatiive □ No effect □ Slightly positive □ Very positive

Q12. Did the review exercises contribute to your understanding of the importance of quality assurance of conceptual models?

□ Nothing □ A little □ Medium □ Much □ Very much

Q13. Did the review exercises contribute to your understanding of the importance of conceptual modelling in IS development?

□ Nothing □ A little □ Medium □ Much □ Very much

PART V: LONG TERM LEARNING

Q14. Do you think the review activities have made you better prepared for later theory courses?

□ Nothing □ A little □ Medium □ Much □ Very much

Q15. Do you think the review activities have made you better prepared for later project courses?

□ Nothing □ A little □ Medium □ Much □ Very much

Q16. Do you think the review activities have made you better prepared for later work-life?

□ Nothing □ A little □ Medium □ Much □ Very much

PART VI: THE REVIEW PROCESS

Q17. How satisfied were you with the feedback from those who reviewed your models?

□ Very dissatisf. □ Slightly dissatisf. □ Medium □ Satisfied □ Very satisfied

Q18. As a reviewer, would you have liked to know whose models you were reviewing?

□ YES □ NO

Q19. As a reviewer, would you have liked to communicate with the modeller, for clarifications about the contents of his/her model?

□ YES □ NO

Q20. As a reviewer, would you have liked to cooperate with others reviewing the same model?

□ YES □ NO

Q21. As a modeller, would you have liked to know who reviewed your model?

□ YES □ NO

Q22. As a modeller, would you have liked to be able to respond to the reviewers' comments to your models?

□ YES □ NO

Q23. How much time did you spend in total (both review exercises) to look at and comment on others' models?

APPROX. NUMBER OF HOURS: _____

Q24. How much time did you spend in total (both review exercises) to look at feedback on your own models (and, possibly, to change your models according to the comments)?

APPROX. NUMBER OF HOURS: _____

Q25. Do you have any other comments or suggestions as to how the review process can be improved?

_____

_____

_____

Information Systems & Computing
Academic Professionals

STATEMENT OF PEER REVIEW INTEGRITY

All papers published in the Journal of Information Systems Education have undergone rigorous peer review. This includes an initial editor screening and double-blind refereeing by three or more expert referees.

ISSN 1055-3096