EXPERIENCES WITH PEER-REVIEW EVALUATION IN COMPUTER SCIENCE COURSES

Mark, Burgess[1], Frode Eika Sandnes[2]

Abstract *¾ Peer-review is commonly associated with the publication of research results in international conference proceedings and journals. Variations on peer-review are also a fairly well-established means of evaluation in humanitarian and arts oriented educational programs where there is a stronger emphasis on written assignments and socio-cultural learning than is the case in typical engineering programmes loyal to the behaviouristic teaching model. At the faculty of Engineering at the Oslo University College in Norway we have been experimenting with various flavors of peer-review as a form of evaluation in two large-class final-year computer science courses. In the first of these courses – computer security, students themselves evaluated each others progress and provided feedback. The peer-review process was aided by an online peer-review system similar to those used at international conferences. In the second of these courses – application development, peer review was employed by students to evaluate the quality of web application implementations based on a set of predefined criteria. These trials were performed manually without the aid of a reviewing tool. This paper describes how the peer-review activities were realised, their effectiveness in terms of reduced workload on the instructor and as a source of constructive feedback for the learner. Further, the article elaborates on how students responded to this style of evaluation.*

Index Terms ¾ Student evaluation, peer-review, large-class teaching, student activity.

INTRODUCTION

Norwegian law currently requires every examination, which counts towards a student grade, to be examined by one internal and one external examiner. That means that, in a small country like Norway, not only do staff have to grade their own exams, they also have to grade at least one other, from an external university. The burden of this process has been excruciating, and extremely costly. Our college currently uses more funds on grading courses than it does in teaching them --- simply to pay external examiners to do a job, which they perform only grudgingly, and which presents them with no academic challenge or job satisfaction. In addition, it has become harder to find qualified examiners, expecially in newer subjects. All of these factors pointed towards the need for a labour saving means of examination in which the *security* for the college and students is at least as good as it is under a traditional examination system.

A CRITIQUE OF EXAM BASED EVALUATION?

Students are exposed to information for a period of time, and then tested on their absorption, recollection or understanding of the material, in a timed monologue. The function of the exam, if not its intent, is to discover what students do *not* know, so that the college can distinguish and attest to `levels of achievement'. The strategy of a written exam is usually to pitch questions at a level of `difficulty' which sorts the `good' students from the `bad' students, and allows the good to excel, forces the bad down the grade scale, and which discourages the ugly from even registering.

There are many quotes in the preceding sentence, because the concepts are somewhat ill-defined. What *is* a good student, or a bad student? What is difficult? Usually, when this is put to the court of opinion, one fends criticism with mention of statistics and Gauss curves of ability. Some teachers even believe that the correct way to grade an exam is to fail a certain fraction of the whole, and base the grades on an approximate Gaussian fit. Of course, this is statistical nonsense, because it is based on the assumption of universally comparable conditions in every exam.

Whether or not it is impartial, it is certainly not unique. Grade levels are usually a political issue. Many teachers feel that giving low grades is a mark of quality – i.e. that by conceding little, they keep standards high. Often this has the opposite effect, however; students quickly realize that there is no point in trying hard, because they will never achieve a good grade.

The final examination is motivated by the desire for student quality control. The belief is that, if the students have acquired the skills and knowledge which are required of them, then they will be able to pass the exam; if they have not, they will not, and the learning institution will not endorse their abilities. The strategy might be called an exclusive one – it purports to categorize students, in a real sense, into difference classes of ability. The trouble is that the integrity of the examination system is open to attack in many forms. Students who know how to pull examiners' strings, can pass with reasonable grades, without really understanding the content of the course. This begs the question, what does the exam actually test?

Another criticism of the final examination is that it tests quality too late, and manages only to fail the bad

[1] Mark Burgess, Faculty of Engineering, Oslo University College, Cort Adelers gate 30, N-0254 Oslo, Norway mark@iu.hio.no
[2] Frode Eika Sandnes, Faculty of Engineering, Oslo University College, Cort Adelers gate 30, N-0254 Oslo, Norway frodes+icee@iu.hio.no

components emerging from the factory. Unless it is part of a carefully thought out programme of feedback, it fails to evaluate what went wrong, or correct the deficiencies before it is too late. It is a potentially wasteful approach to quality management. No attempt is made, by exam evaluation, to prevent failure; nor is it clear what failure means, since the process of learning is divorced from the process of examination. Students have no compelling reason to do any work until immediately before the examination period. Failure can occur for many reasons: poor teaching, laziness of teacher or student, distractions from other courses, badly worded exam questions, and so on. Exams must therefore be allied with some other form of *study management*, in order to create a successful course.

A general dissatisfaction with the written exam, (including the difficulty of inventing suitable questions in several subjects), led therefore to the desire to try to address some of these points in a new scheme. An alternative approach to student evaluation would be to use a scheme of continual evaluation . This is not new in some subject areas, but it is not an approach which has been widely used in scientific or technological disciplines [5]. With a continuous appraisal, evaluation would be not only of student progress, but of the teaching process itself; it might prevent total failure, and avoid the wasted resources involved in having to re-take courses. Rather than trying to find faults in the students at the last minute, one wipes the slate and begins with a new philosophy: the aim becomes to get as many students as possible to work and achieve, during the course. Any such scheme has to be able to address all of the causes of failure, including staff and student deficiencies, human error, and so on.

SECURITY ISSUES

Any system in which points are awarded, or students receive some kind of reward (payoff), is subject to attack either by malicious or incidental factors. Let us mention a few of the ways in which the tenets of security apply to the evaluation process.

- ***Trust*:** The fundamental issue in any security system is where one places one's trust; it is about deciding what is an acceptable risk. For example, staff might trust students never to cheat, or staff might only trust students not to cheat in a supervised room (with an exam invigilator present). Conversely, students might not trust the course teacher to grade their papers correctly, or to give them a fair hearing.
- ***Reliability***: The reliability of the examination procedure must be secured against both malicious exploitation and accidental error. If machine can be made to perform the grading [1, 3, 4, 6, 7], then clearly the only source of error would be a systematic error, perhaps from an error in programming of the system itself.

Some problems have no right or wrong answer, but simply need to be graded on experience and reasoned opinion. This means that humans have to be involved, and redundancy is required to ensure that individual emotional and personal aspects do not have a significant impact on the result. This means that multiple examiners, with clear guidelines, are required to avoid fatigue and differences of understanding.

An aspect which current examination systems do not address is the reliability of the learning process itself. Students cannot be guaranteed a fair grade, unless the teaching process itself was adequate – and this issue disjointed from the examination process.

- ***Integrity***: Integrity concerns the ability to transmit information, or intent, without alteration or error. Integrity of evaluation information applies both to the problems posed to students and in the collection their replies. Once an exam question or work problem has been posed, is it typed correctly, is the paper reproduced correctly, are any pages missing?

Even if no error has occurred in reproduction or dissemination of the information, it could be that the question was ambiguously formulated, so that the intention was not correctly disseminated. Similarly, access to information required to answer the problem needs to be assured: was the material actually covered in the lectures?

The integrity of an evaluation scheme can be attacked by external political pressure. This is seen very easily by noting that it is in every College's best interests to give every student a high grade, regardless of what they know. Norwegian Universities are now being `encouraged' to do this, by a government education policy which awards funds based on the numbers of students who pass exams. No penalties are exacted for low quality however; thus, as a problem in pure economics, learning institutions would do best by simply handing students their certificates when they arrive, and by saving money on teaching and examining. The state places its trust in the integrity of the academic institutions, but at the same time places an incentive (or even pressure) on them to cheat.

- ***Authenticity and identity***: Students need to trust the authenticity of the exam paper, or the problems they are to answer. It would be unacceptable for a malicious party to replace the actual exam with a fake exam, or an exam to which the students already had the written solutions. Similarly, the examiners need to know that the student whose name is on the resulting work, actually did that work.

Correctly identifying the author of an examination paper is a subtle task. In the security sense, one can visually inspect the student ID of a student who shows up for an examination (though ID can be forged). Similarly, one can forge electronic credentials relatively easily. In spite of the dangers to themselves, students regularly swap passwords and loan accounts to their classmates. Thus, when an assignment is submitted without physical supervision (e.g. electronically), there is no guarantee that the person whose name is registered by the receiver is the author of the work. In a written examination, students regularly memorize

passages and methods written by others however -- it just requires a little more concentration to achieve.

PEER-REVIEW FOR QUALITY ASSURANCE

Peer-review is a well-known technique that is widely used for assuring quality when publishing research results in international journals and conference proceedings. Manuscripts submitted to the conference program chair or journal editor are distributed to anonymous reviewers. These peers are other researchers from academia and industry with a similar competence and level of "expertise" working on similar problems. The number of referees varies from one up to five, but three referees are most common. Each referee studies the manuscript according to a set of guidelines, makes comments, rates the manuscript and returns the manuscript to the coordinator or editor. The program committee or journal board of editors use the referee reports to make a decision on whether to include the paper in the journal or conference proceedings and notify the authors of the decision. The authors are given a partial or full copy of each referee report. Occasionally a manuscript undergoes several iterations of peer review before it is finally accepted or rejected. This is common practice in international journals and prestigious conferences. The peer-review process aids both in quality and managerial issues such as:

- Load balancing – the peer review process disperses the workload associated with reviewing a large number of manuscripts. Instead of a program chair, editor or a small number of panel members having to thoroughly read through every single submission they can focus their effort on more high-level issues. The tedious job of carefully reviewing a manuscript is left to a large group of reviewers – each responsible for reviewing a small number of papers. The workload is evenly distributed.
- Reliability and redundancy – by involving a large number of people in a project it is nearly impossible to avoid some yield, i.e. certain percentage of the involved individuals will fail to perform their assigned tasks. By assigning several reviewers to each manuscript one can afford some yield.
- Coherence and bias – "Everyone has the right to a fair trial" - certain structured, quantifiable and undisputable criteria can be used when evaluating a manuscript; however, there are many aspects of the evaluation process that cannot be quantified neutrally. Assessment is based on the reviewers own experiences, academic interests, insight, mood and intuition. There is thus plenty of room for bias – both negative and positive bias. Multiple referee reports assimilated from multiple reviewers are more likely to found the basis for a more neutral decision. Imagine three reviews, two positive and one negative – then, it is likely that the majority is right and that the negative review is biased. Each reviewer has no knowledge of the other reviewers.
- Quality – journals and conferences both strive for a reputation of high quality and excellence. The refereeing process helps to achieve this by several means. First, poor submissions are detected and filtered. Any call for paper will attract a certain element of outliers that do not fit naturally into the specialised community. Second, everyone makes mistakes. An otherwise excellent manuscript can be ruined by banal mistakes. The peer-review process is a particularly efficient strategy for detecting and eliminating such mistakes, provided the peers can be trusted.

- **Learning** – the peer-review process provides a bi-directional communication channel. First, the reviewers may be inspired by a paper that they perhaps otherwise would not have bothered to read. Secondly, the author gets expert advice and comments and suggestions reflecting a different angle and different knowledge. This conglomeration of knowledge is particularly important for the advancement of scientific research.

Many of these characteristics are desirable in the context of teaching and are directly applicable. Further, the idea of using peer-review evaluation in teaching is well established within certain disciplines – in particular, arts subjects such as language learning, literature, pedagogy etc, although the form of peer-review may be less rigorous than the one used in scientific reporting. To the best of our knowledge, this form of evaluation is rarely used in engineering education, despite the technique being quite familiar to a broad cross-section of the faculty members conducting research.

EXPERIENCES FROM THE COURSE "COMPUTER SECURITY"

In our trial we used group-based peer review. Groups of three or four students would work together on a project, and each group would submit an assignment together. The reports would then be distributed anonymously to three or four other groups who would grade the work, according to a set of common guidelines. In this way, no group had more than three or four other reports to grade, and no group was reliant on the mercy or competence of a single reviewer. This procedure scales without limit. No student group reviewed their own work and no group reviewed the same work more than once; an automated system solves this problem easily.

When work is carried out in groups, there is always the question of who did what, and whether the contributions were evenly spread throughout the group. At Oslo, we have seen every possible combination of group compositions: students seek to work with peers whom they like, rather than judging ability.

No grade-differentiation was attempted between the different students in a group. This was considered to be too difficult a technical problem to solve, in a fair way, and differences were best addressed by additional personal tests of the other types, e.g. by posing MC questions about the work they had submitted. However, this admits the possibility of error, if one thinks in terms of a traditional grading strategy; but mixed groups can also have a positive impact in disseminating knowledge from `clever' to `less clever' students. We know of no studies which indicate how this collaborative process might work, in general. Our strategy here is in line with the revised philosophy of being less interested in judging student merit, as in encouraging learning. The belief here was that the final results would likely be better, for this choice of strategy, and whatever the reason, we were not disappointed.

When submitting group work, students were made to authenticate their work together, by typing in their normal system login name and password, in order to `sign' the work with a simple electronic signature. By insisting that this be done simultaneously from the same terminal, we hoped to force the students to meet physical and resolve any differences before submitting. The login box, for uploading an assignment, asked users to enter their user names and passwords (as a primitive digital signature). This was used as a mutual authentication of students by one another, and an affirmation of who deserved credit for the project. Our experience with project work over many years has led to only a tiny number of cases in which students revoked one anothers' membership in collaborative groups.

An advantage of peer review, over a single examiner, is that it is possible for the examiner to know exactly what the students taking the course should know (because the examiners took the course themselves). No one has more intimate knowledge of the course details than the students themselves. By ensuring that every report has at least three reviews, one hopes to mitigate implausible grades. Moreover, with the new philosophy of using grades to motivate achievement, rather than to catalogue failure, the precise grade is of less interest than what the student ends up learning. In our tests, the students were only allowed to set grades in a few categories, not to set arbitrary grades.

Peer reviewed and examined work

An advantage with the peer review scheme is that students not only write their own assignments, but have get up to speed themselves in order to grade each others'. Why should students bother to grade each others work? By giving students a substantial `grade reward' on submitting their reviews, there is an incentive for them to take this work seriously. The threat of a penalty is also maintained, if they abuse the privilege, or do an incompetent job. Our experience, however, is that students are flattered by their additional responsibility, and take the job seriously. The peer review process has other advantages too; it

- exposes students to each others' work, showing them where they lie in relation to others,
- forces students to look at the problems from other groups' perspectives,
- encourages them to reevaluate their own performance, based on what they have seen.
- gives them a glimpse of how evaluation works in the `real world'.

Students at Oslo reported that they learned a lot from this procedure, both in terms of hammering home the content of the assignment, and also in terms of approaches to working. Some expressed concern however which indicated less than total trust in their peers. Several students expressed the viewpoint that they would likely receive a fairer grade with a traditional examiner. The mistrust was mostly unjustified.

Fatigue of the examiner should not be underestimated as a source of unreliable grading, in regular examinations. In the peer-review trials, students expressed a greater trust in an external examiner than in their student peers, but failed to see that the attention to detail, or `personal service' which they received from their peers was, by and large, far in excess of that which an external examiner would have mustered.

External examiners vary as much as the students, from kindly benefactors to self-righteous critics. Once again, we have seen no justification for the view that anything is lost by omitting external examiners. The contention of our trial was that, on the contrary, peer review can be fairer than using a small number of external reviewers, provided the class is large enough, because it randomizes the choice of examiners for each submission. Especially when it is difficult to find qualified examiners, in special subject areas.

On average, students can expect to receive an average quality level of review. If one commits the sin of assuming that student aptitude is distributed in a Gaussian fashion about some mean value, then choosing a random sample of three or more reviewers per submission is more likely to result in an even grading of problems, than is choosing a single external reviewer, whose knowledge of the course and its materials is quite unpredictable. Moreover, by making groups grade other groups, there is --- at least, in principle --- an addition local averaging over the group opinions, though sometimes a single group member will dominate group opinion. At Oslo University College, the students have experience of, and are good at collaborating in groups, though this strategy might not be as successful in other settings.

Results

The results obtained from the trial indicate that it has been highly successful. Using a course evaluation, and interviews with students, including a final control project, graded by an external examiner, we could see how well the peer review worked, and how well the grades reflected student ability.

The strategy of using the grade to motivate work, rather than to gauge failure, had no apparently negative

consequences. Grades were generally higher than in previous years, and the actual level of understanding achieved by the students was also gauged to be higher. Although this is not a fully significant finding, it is indicative of the success of the strategy.

Even the weaker students achieved respectable grades, at the high end of the scale. Of course, only a long term follow-up study in a job-unrelated topic (e.g. basic physics or maths) could tell how well their knowledge remains over time, compared to traditional methods.

The key point in all of this is that, under a scheme of continual evaluation, students must receive a *meaningful* credit for everything they do. Students have a keen sense of fairness and are quick to question and even complain if they do not see justice being done --- the right amount of credit for a given task must be given to make it worth their while. Although some teachers will find this cynical, it is a regrettable development of our market driven society and is to be be ignored only at out peril.

In summary, there are two strategies for using grades: as a *certificate of competence*, and as a *motivating reward*. There is insufficient evidence to conclude which of these strategies is appropriate or when, but such evidence will never emerge unless trials are done.

EXPERIENCES FROM THE COURSE "APPLICATION DEVELOPMENT"

The course application development is targeted at final year computer science students and aims to cover current state-of-the-art techniques and technology used in industry to develop e-commerce, business to business and corporate applications. The course addresses three tier architectures with emphasis on web based interfaces (Servlets and Java Server Pages) and thin clients such as mobile phones (WML, i-mode), relational databases and XML. Students have to complete three compulsory assignments during the course and these projects are usually carried out in teams comprising of up to four individuals. The course is an optional module and approximately 100 students enrolled.

Peer-Review Evaluation Applied

Peer review was applied to the first of the three compulsory assignments. The task was to design and implement a web based survey system enabling an administrator to create and modify a questionnaire using a web interface, the deployment of the web-based questionnaire for a survey on the Internet and a web interface for examining survey statistics. The application had to be implemented as a Java Servlet and three weeks were set aside to complete the project.

Students were instructed to adhere strictly to the announced deadline – failure to comply with the deadline would result in an exclusion from the evaluation process. Students were asked to submit their solution attached to an email sent to the instructor, where the email message listed the group members and the subject line of the message contained the predefined token "apput1". Incoming messages were placed into a designated folder, separating them from unrelated correspondence. After the deadline the instructor examined all the submission making a list of all students, arranged into groups. The instructor assigned three other individuals from different groups to each member of each group. No individual reviewed his or her own work and no individual reviewed more than one submission from the same group. No two people reviewed each others work. The refereeing list was compiled manually using an spreadsheet and it took about two hours to complete. The list of assigned referees was published on the course web page. In the following lecture students were informed of the peer review process and told to look up their name on the list to see who their referees were and who's work they had to evaluate. Further, they were told to send their solution to the other referees on the list and expect to receive a specimen from three other individuals to evaluate. They were given one week to complete the exchange of projects and one week to complete the peer review. Reviews were returned via email to the author and to the course instructor. The criteria for the evaluation fell into three classes – installation and deployment, user interface design and technical finesse. Installation and deployment entail the following:

- **Installation** – the applications were packed in an archive ready for deployment. The students assessed the installation procedure for the application and reported on potential problems.
- **Deployment** – how easy was it to deploy the application? Did it run immediately? Was any tweaking necessary?

These criteria are more naturally addressed using textual information. However, the user interface evaluation was done numerically on a scale from 1 to 6 where 1 is the best . Textual comments were also encouraged. User interface design was evaluated using the following subcategories that are commonplace in user interface usability testing [2]:

- General usability – how easy is it to use the application in general?
- Efficiency and ergonomics – how efficient is it to use the application?
- Navigation and orientation – is it easy to navigate the application, i.e. to locate the various functions?
- Help and assistance – is the application self-explanatory? Is there instructions and help available?
- Learning – how long does it take to learn to use the application?
- Security and robustness – how difficult is it to make mistakes? I.e. accidentally erasing all the entries etc.
- Aesthetics – is the application well presented?
- Functionality – does the application adhere to the specification?

Finally, **technical finesse** was evaluated by examining the source code - to see if it is well laid out and commented, and to discover impressing features or finesse.

Experiences and observations

The vast number of email messages interfered with the instructors daily routine. Initially, about 40 email submissions were received as only one email message was sent per group; this is a manageable quantity. However, during peer-review in excess of 250 emails were received, as nearly each enrolled student sent three referee reports to each peer and to the instructor. These messages were also sent over a relatively short time interval at set times of the day when students did not have lectures etc.

Secondly, a significant, but manageable, number of communications were received regarding peers that did not respond, or peers email address that could not be found – although these are published on the faculty website. Students were told to ignore non-responding peers and were directed to the URL of the faculty email address list for missing email addresses.

The automatic mail filtering configuration also caused problems. A couple of students did not follow the instructions providing the email subject line – consequently these emails had to be manually moved to the designated folder. More of a problem was it that some students ordinary email queries were labelled with the same token in the subject header as the peer review emails. These messages were not immediately discovered since the mails in the designated folder were inspected in batch at specific times. Consequently, some students did not get the required assistance.

Informal interviews were used to evaluate the effectiveness and usefulness of the peer-review activity. The interviews revealed a couple of patterns. Many students claimed they found the review process too time-consuming – which they felt unfair as they had already devoted a significant amount of time to the course and they were pressed for time due to other courses.

TABLE I:
DISTRIBUTION OF RESULTS FROM AN END OF TERM EXAM AND STUDENTS PEER-REVIEW OF THE FIRST COMPULSORY ASSIGNMENT.

Result	Exam	Peer-review
1.0-1.9	31.3 %	27.5 %
2.0-2.9	43.4 %	47.3 %
3.0-3.9	18.3 %	22.2 %
4.0	0.0 %	0.0 %
Fail	6.0 %	3.0 %

Another common trend was that students had difficulty in installing and deploying the applications of others. Hence, many students never got beyond the installation step which is a prerequisite for evaluating the application. As this was the first compulsory exercise the correct packaging of applications was a skill not yet acquired by everyone. This skill improved with the second and especially the third compulsory exercises that followed.

An opinion expressed throughout the interviews was that students generally were unhappy with their own work. This was their first time making web applications, and there was an overwhelming amount of new impressions to digest – and several students were pressed for time. Consequently, they felt uncomfortable submitting incomplete or unfinished projects. Thus, many of the issues raised in the feedback reports were already known to the authors. An implicit expectation amongst the students was to learn something new though the peer review.

Table I shows the distribution of marks issued by the students in the peer-review process (column two), and the results of the end-of-term exam. The distributions match quite well. However, note that the data do not show the correlation between the results for individual students.

Finally, many students expressed the view that they enjoyed this form of evaluation, and that they learned from seeing other groups' realisation of the given specification. However, some found that the activity appeared haphazard, unplanned and badly organised.

CONCLUSIONS

Our tests have shown that peer-review can be used sucessfully in computer science related subjects. The tests have revealed that it is important to issue very clear and structured guidelines to the students on how to conduct the peer-review. Futher, it is advicable to employ computer assisted tools in the review process to avoid misunderstandings and lessen the workload.

REFERENCES

[1] Alexander, S., O'Reilly, U., Sweeney, P. and McAllister G., "Utilizing automated assessment for large student cohorts", ENGINEERING EDUCATION AND RESEARCH – 2001: A Chronicle of Worldwide Innovations, iNEER and Begell House Publishers, 2002.

[2] Cato, J., "User-Centered Web Design", Addison Wesley, pages 191-221, 2001.

[3] Cyberchair, "An Online Paper Submission and Reviewing System with proceedings preparation support", www.cyberchair.org.

[4] da Silva, J.G.S. de Almeida, N.N Santiago, R.A.. da Silva H.V Santos, A.R. Duarte, "*Development of a Graphical and Interactive Electronic Workbook Based On The Homepage Concept*", Proceedings of the ICEE 2001 Conference, 2001

[5] Khambadkone, A, "Criterion-based continuous assessment towards in-depth and student-centered learning: a case study", ENGINEERING EDUCATION AND RESEARCH – 2001: A Chronicle of Worldwide Innovations, iNEER and Begell House Publishers, 2002.

[6] Sehcrest, J."*Interquest and Questwriter, Experiences in Online Classes*", Teaching in Community Colleges Online Conference (TCCOC), 1998.

[7] Wyer, M. and Eisenbach S., *LEXIS: an exam invigilation system*, Proceedings of the Fiftteenth Systems Administration Conference (LISA XV) (USENIX Association: Berkeley, CA, p199, 2001.