SUMMATIVE PEER ASSESSMENT USING ‘TURNITIN’ AND A LARGE COHORT OF STUDENTS: A CASE STUDY

Silvester Draaijer and Patris van Boxel

Summative Peer Assessment Using 'Turnitin' and a Large Cohort of Students: A Case Study

Silvester Draaijer and Patris van Boxel
Centre for Educational Training
Assessment and Research (CETAR)
Onderwijscentrum VU
Vrije Universiteit Amsterdam
De Boelelaan 1105
081 HV Amsterdam
+31 20 598 54 79
s.draaijer@ond.vu.nl
p.vanboxel@ond.vu.nl

Abstract
At the Vrije Universiteit of Amsterdam, the use of peer assessment is increasingly being considered by lecturers that want to give their traditional lecture-based courses a more active learning component. Prins et al. (2005) point out that peer assessment can be very well integrated in such courses and research shows that formative peer assessment results in an increased understanding of the learning content, the development of assessment skills and a reflection on one's own learning performance (Hamer, Kwong et al. 2005; Prins, Sluijsmans et al. 2005). However, the validity and reliability of peer-generated marks is still under debate (Cho and Schunn 2003). In order to support peer assessment in courses with large cohorts of students, computer support can be regarded as a necessity to manage the whole process of assignment submission and grading.
This paper describes a case study on the use of a commercial peer assessment application for summative peer assessment. It describes the course set up, the use of the software and use of the generated marks for summative purposes. The study shows that the system is easy to use for both instructors and students and that it can support large cohorts of students despite some technical problems. The students are very positive about the benefits of peer assessment for their own learning, but they have a low confidence in peer assessment for summative purposes, despite considerable efforts to motivate the students and to build in measures to increase the grade reliability and validity.

Introduction

Peer assessment
The shift of learning and teaching from a knowledge centred activity to a competence based activity calls for new forms of assessment that go beyond traditional knowledge testing. Amongst others, Dierick (2001) and Van den Elsen (2005) describe how an 'assessment culture' is emerging in which several assessment formats are combined to assess a broad range of knowledge, skills and competencies throughout the curriculum and learning process.

One characteristic of the assessment culture is that appraisal no longer solely takes place at the end of a unit of learning, but becomes an integral part of the learning unit. Moreover, it is expected from students that they play an active role in the assessment process. Peer assessment is a form of assessment that requires such active student engagement. Formative benefits of peer assessment include an increased understanding of the learning content, the development of assessment skills and an reflection on one's own learning performance (Hamer, Kwong et al. 2005; Prins, Sluijsmans et al. 2005). In addition to learning benefits, time saving is also often given as a pragmatic reason to introduce peer assessment, especially when it allows for more opportunities for personal feedback in a large student class.

Design of Peer Assessment Assignments

The use of peer assessment assignments must be closely related to the regular course material and preferably integrated in the course assignments (Prins, Sluijsmans et al. 2005) Langan (2003) and Keatley (2004) point out that the success of peer assessment depends greatly on how the process is set-up and subsequently managed. Keatly summarizes the guidelines that several authors have provided for the management of peer-assessment (e.g. Stefani 1994; Topping 1998; Race 1999; Magin and Helmore 2001; Ballantyne, Hughes et al. 2002; Prins, Sluijsmans et al. 2005):

- explanation of the benefits and rationale of peer assessment
- clear procedural guidelines
- clear assessment criteria
- access to concrete examples of assessed work, where possible include practice sessions using the assessment criteria to mark "good", "average" and "poor" exemplars of student work
- a complaints or review procedure so that peer awarded marks can be discussed/challenged;
- some form of feedback to students to confirm that peer marks are reliable and similar to that of their tutors.

Perhaps these guidelines have contributed to the growing evidence that students are able to assess each other (e.g. Topping 1998; Hughes 2001). Furthermore, standardised assessment criteria, multiple assessors and detailed instruction are believed to significantly reduce biased marking. Students also take more care when they themselves are being evaluated on their marking ability (Ballantyne, Hughes et al. 2002).

Summative Assessment

Although there may be little debate about positive reasons to develop skills associated with peer assessment in learners, the validity and reliability of peer-generated marks is still under debate (Cho and Schunn 2003). Dierick (2001) argues that the new assessment culture also challenges our traditional views on these concepts.

According to Dierick, grading validity can be increased through transparency of the assessment procedure, authenticity of tasks and access to criteria for evaluating performance. Hamer et al. (2005) specifies the latter by recommending the use of scoring rubrics, a descriptive scoring scheme that guides to reviewer in assessing various aspects of work. Reliability can be computed by comparing ratings between

individual student evaluators. Hamer et al. (2005) and Davies (2005) also contributed to this by developing grading algorithms that identify the grading quality of students.

Online Peer Assessment

In the last five years, dedicated software applications have emerged that support the online organisation of peer assessment assignments for both small to large cohorts of students (Davies 2000; Chapman and Fiore 2001; Davies 2002; Parsons, Handy et al. 2003; Volder 2005). Online peer assessment can ensure the anonymity of both reviewer and student being reviewed, which may contribute to more objective scoring (Ballantyne, Hughes et al. 2002). It also allows for easy submission of assignments and redistribution for review.

Case Study

Case study set up and aims

At the Vrije Universiteit Amsterdam, the use of peer assessment is increasingly being considered by lecturers that want to give their traditional lecture-based courses a more active learning component. Also, the University's Centre for Educational Training, Assessment and Research (CETAR) has a keen interest in promoting and supporting learning methods which stimulate self-reflection and collaborative learning, such as peer assessment

In the academic year 2005-2006, a pilot was set up to introduce a number of online peer assessment assignments in a third year Marketing course (Consumer Behavior) at the Faculty of Economics. Two hundred and fifty students enrolled in the course.

The lecturer that participated in the pilot was very positive about the potential learning benefit of peer assessment and the possibility to use it with a large group of students through a web-based system. He developed five peer assessment assignments and decided to use the outcomes partly for summative purposes (40% of the final grade derived from peer grades; 60% from the written examination). To increase student commitment, a small proportion of the grade was derived from the student's accuracy as an assessor.

Keeping in mind the large number of assignments that would be submitted, the even larger number of peer reviews this would generate (>5000) and the potential problems to use peer review summatively, it was important that the selected application:

- was easy and fast to use;
- supported a process of double blind review (anonymising both students and reviewers);
- could check for plagiarism;
- supported student grading on the basis of predetermined assessment criteria;
- supported free text feedback.

On the basis of those criteria, the web-based application Turnitin was selected.

The aims of the pilot were to seek answers to the following questions:

- Can Turnitin support an online peer review process involving a large cohort of students?
- How easy is it to use such a system for both the instructor and the students?

- Does the use of a double blind review and a grading algorithm which identifies poor graders and diminishes their contribution in favour of accurate reviewers, increase students' confidence in the reliability of peer assessment?
- Can peer assessment be used for summative purposes? What are the arguments to support this or to abandon this?

Peer review procedure
The Marketing course was designed as follows:

- The course had a duration of 6 weeks, each week, a lecture was give to the whole student group.
- Every week, the students had to complete an assignment. These assignments had to be submitted on a Tuesday before 17.00 hours.
- After submission, each assignment had to be assessed by 5 other students. Those students were randomly and anonymously assigned to the assignment and the identity of the authors was also not disclosed (resulting therefore in a double blind review).
- The peer reviews had to be completed before Thursdays 17.00 hours.
- The students had to grade each essay via ten questions on a 1 to 5 scale and 1 open question.
- Final grading and calibration was completed and results reported back to the students by the following Tuesday
- In total, about a thousand essays were submitted and a total of around five thousand scores were assigned.

Assignment description
Students had to complete five assignments in total. Each assignment had to be worked out into 500 to 800 word text document. After submission of each assignment, they had to review the work of 4 or 5 fellow students.

Assignment support materials and information
The students received detailed information on the peer assessment procedure. Amongst others, the following guidelines were provided:

- A clear overview of deadlines (for submission of assignments and peer review assignments)
- All assignments were to be submitted anonymously (no names or student numbers in the assignment or filenames).
- The students were informed that 40% of their final grade for the course would be derived from the grades they were awarded by their fellow students. They were also informed about the use of the calibration process to establish a reliable assessment procedure and to determine their quality as an assessor. This process is based on the work of Hamer (2005) and explained as follows: suppose that you are graded by student A and by student B. Now suppose from benchmark data, it is established that student A is a very accurate grader; then the grade awarded by student A will count for 100%. On the other hand, when student B shows to be a very non-accurate grader, his score will only count for say 8%.

Peer assessment criteria
The students had to assess their peers on the basis of 10 closed questions and one open question. They were given performance scoring rubrics for the closed

questions. Using these rubrics, students could compare the performance of their peers to a set of predetermined standards by the lecturer. This was also believed to increase the validity of the peer review process.

Score calibration
The calibration process on basis of the procedure as proposed by Hamer (2005) was quite labour intensive to execute. The first step that had to be taken with each assignment was to extract all data from Turnitin. Because the system does not support a direct download of data, a number of manual actions had to be undertaken. The next step was to perform the calibration calculations of all the final marks for the assignment within MS-Excel.

In the original course setup, it was envisaged that the lecturer would also mark a number of assignments to allow for further score calibration and assess the validity of the allocated grades. However, due to major time constraints, instructor marking and calibration was not carried out.

System Performance
During the first week, Turnitin turned out to be very unstable and unreliable. The assignment of the first week did not work out at all. The following problems were reported.

- *Speed* (the system seems slow at times, gives a time out, data entry is broken off by the system and the user returns to the Turnitin home page). This seemed to be particularly the case shortly before deadlines when many students were working on the peer review assignment simultaneously.
- *Double entries*. Students got exactly five essays to mark, but sometimes got one extra essay to mark which always turned out to be a duplicate of one of the five already assigned essays. Also, when students were thrown out of the system (see above), they might find they could not open the broken off assignment, but also encountered a new entry for that assignment.
- *Unable to open an essay for review*. Sometimes students were unable to open one of the five assigned essays for review. Strangely enough, if the lecturer logged in and tried to open it as supervisor, he often had more luck - though not always!
- *System errors*. In week 1 the answers to all the open questions were obliterated by Turnitin, and in week 4, the anwers to the open questions were temporarily unavailable.

An intensive communication between the instructor, the system administrator and Turnitin ensued to determine and resolve the problems. It was decided to regard the first assignment as a ‘trial’ and not to use the scores of this week for summative purposes.

The assignments of the next four weeks also encountered technical problems, but they were not so disruptive as the problems of the first week. During the next weeks, close attention was paid to trace and resolve problems as soon as possible.

Evaluation of the Case Study

Data Collection

In order to evaluate the pilot, data were collected about:

- the attitude of the students towards peer assessment as learning method and assessment instrument and
- the student's and lecturer's satisfaction with the use of the Turnitin system

These data were obtained via a questionnaire and a focus group interview:

- After completion of the course, students were asked to complete a questionnaire with 20 statements. They could reflect on a 1 to 5 scale (1 meaning 'I disagree strongly with the statement' and 5 being 'I agree strongly with the statement') . They were also asked to describe on their own words what value peer review had for their own learning.
- Six students took part in a focus group interview. This was led by the CETAR educational advisor of the project, the lecturer was not present so student could speak more freely.

The lecturer evaluated the software and course setup via email correspondence with the CETAR educational advisor.

An analysis of the score allocation and distribution over the five week period is due to be carried out. This might give further insight into whether students' scoring strategies and abilities change over a series of peer assignments.

Appraisal of the system by the lecturer

The lecturer reported that the planning and execution of the peer assessment process went mostly as planned. The Turnitin interface was quite straightforward to use, and particularly the option to quickly create a double blind review process for a few hundred students was crucial to the success of the pilot. The instructor was pleased with the strong integration of assignment submissions, marking and feedback in a single interface. The instructor would like the option to also be able to deploy other than only 1 to 5 Likert scale rubrics.

Although he expected to put in a reasonable effort to set up and manage the peer review process, he still put more time than he had anticipated. However, organizing a series of five peer assignments for 250 students would not have been possible to realize at all without a system such as Turnitin

Appraisal by the students

The data that were collected via the questionnaire are partly summarized in Table 1. They describe students' opinions about Turnitin and the perceived learning benefit of peer assessment.

		1	2	3	4	5	N	Àv	SD
Q1	The peer review software (Turnitin) is easy to use	0	6	6	101	49	162	4.19	.67
Q2	Peer assessment is a good method to work with learning content	9	28	49	62	11	159	3.24	1.01
Q3	I learned a lot from comparing my own answers with answers of my peers	5	30	55	62	7	159	3.23	.91
Q4	I learned a lot from writing commentary on work of my peers	6	37	63	48	5	159	3.06	.90
Q5	I learned a lot from reading commentary from peers on my work	12	40	60	43	3	157	2.90	.95

Table 1 Overview of students' opinions about the learning experience of peer assessment and the system

The data show that students found it easy to work with the Turnitin software (av 4.19), despite the reported technical problems.

In this set up, in which the emphasis was on student's giving scores instead of written feedback, students were particularly positive about the learning benefits derived from reading other students' work (av. 3.23). They were less positive about the benefit of writing comments about work of others and comments from their peers (av. 3.06 and av. 2.90). Still, the data give evidence to the fact that students have a positive attitude towards peer assessment as a learning method (av. 3.24). Partly this was due to the nature of the assignments which asked students to find examples to illustrate theory or to design creative strategies to tackle specific marketing problems. Comparing their own work with work of others, gave them both means to reflect on their own (quality of) work and to get a widened perspective of the learning content.

The open question about the learning value of peer assessment yielded a large number of comments which enforced this finding as illustrated in Box 1.

"Because you see work from others you get a clear and relevant overview, you know your position in relation to others."

"You learn a lot from answers from other students. Often I thought: "I hadn't looked at it in this way"."

"Assessing others means you have to give arguments, and delve into the course content and literature. You play with the content, this means you remember it better."

Box 1 Example student responses towards learning from others through peer assessment

Students' opinions about the use of peer assessment for summative purposes are summarized in Table 2.

		1	2	3	4	5	N	Àv	SD
Q6	Peer assessment is a good method to determine my grade	30	50	60	21	0	161	2.45	.94
Q7	I have confidence that the correct grade for my case is determined by giving different weights to assessors.	22	45	56	35	3	161	2.70	1.02
Q8	The scores I received from peers were generally adequate	15	38	70	36	1	160	2.81	.91

Table 2 Overview of the students' opinion about the use of peer assessment for summative purposes

The data show that the students give a low appraisal of peer assessment in terms of its suitability for summative use (av. 2.45). This was despite the genuine effort of the instructor to guide, motivate and inform the students and to build in measures to increase the grade reliability and validity. Students showed rather low confidence, both in their peers as assessors (av 2.81), and in methods used to calibrate scores and diminish the contribution of poor graders in favour of more accurate reviewers (av. 2.70).

Some questionnaire comments from students suggest that the appreciation of peer grading is particularly negatively affected when:

- score from peers diverge (for example: the same assignment receiving a score of 2 and a score of 5 from different students) and
- written feedback does not seem to be in line with the scores on the closed questions (for example: a peer reviewer states in his feedback that he found the assignment very good, but the score on the closed questions resulted in an average of 3).

Despite the low confidence that students have in grade reliability, the instructor hardly received complaints from individual students about the peer grades they were assigned (less than ten assignments over a total of thousand assignments were submitted for remarking by the lecturer).

The focus group interview produced similar findings to the questionnaire, but gave also additional information: this group included some non native Dutch speakers, who claimed they were repeatedly graded very low, irrespective of the actual quality of their work. It is suspected that the double blind review process is partly responsible for this. A reviewer does not know whether a 'poor' essay is caused by a lack of interest and motivation or caused by a limited seizure of the language. This lack of context about the author is therefore a potential disadvantage when building anonymity into the peer review process.

Conclusions and Discussion

Both lecturer and the students found the Turnitin software easy to work with. Despite initial technical errors, it was adequate in supporting an anonymous peer review process for a large number of users.

The study shows that the software is able to support the process of peer assessment for large groups of students. However, the system must be able to cope with peak-loads, in particular when all students simultaneously upload, download and interact with the system during deadline situations. Although it is crucial in supporting the process of assignment distribution, it does not save time as such for the lecturer, as the didactical setup and management of the process are still labour intensive.

In the case study, peer assessment was used for summative purposes. It was hoped that the assignment design (double blind review, use of scoring rubrics, a large number of assessors and grade calibration) and plenty of guidelines would make students more comfortable with this form of assessment. The evaluation showed that the confidence in its reliability still remained low. Students were on the other hand very positive about peer assessment as learning method, particularly the insight it gave in their own performance and on the learning content as a whole.

The lecturer unfortunately did not have sufficient time to join in the marking process and so his scores could not be used for calibration purposes. The evaluation did not address whether students would be more confident in the summative use of peer assessment if the lecturer would play a role in the scoring process. It was therefore also not possible to determine the (increase) in reliability and validity of the scoring process in relation to a set up with less or more criteria, or other calibration procedures.

The double blind review process may contribute to more objective and reliable scoring, but leaves non-native speakers vulnerable to receiving lower scores. It is important that during the discussion of criteria with students, this issue is addressed sufficiently.

In order to achieve balance between a summative peer assessment setup which students not only find educationally valuable, but which also gives them confidence in their peers as assessors, further investigation into the conditions under which this can be achieved, is therefore necessary. Davies's (2005) approach, that allows for an in between feedback-loop between author and peer-marker, before a final mark is assigned to an essay, should be regarded as an extra reinforcement of a more objective and reliable scoring. This approach however is not supported by the Turnitin system.

Given the above conclusions, the following improvements should be implemented in the Turnitin system to accommodate for required technical and didactical issues:

- To raise the system performance to accommodate for peak loads;
- The option to easily export data to MS Excel or SPSS;
- Preferably more variation in grading scales (currently only 1 to 5 scale);
- Functionality to perform different types of calibration on the scores online;
- Functionality to let students review themselves also and compare this with the other reviewers;
- Functionality to allow for feedback and revision loops between student and reviewers.

References

Ballantyne, R., K. Hughes, et al. (2002). "Developing procedures for implementing peer assessment in large classes using an action research process." Assessment and Evaluation in Higher Education **27**(5): 427-441.

Chapman, O. L. and M. A. Fiore (2001). The White Paper: A Description of CPR. A Writing and Critical Thinking Instructional Tool. Los Angeles.

Cho, K. and C. D. Schunn (2003). Validity and Reliability of Peer Assessments with a Missing Data Estimation Technique. Proceedings of ED-Media 2003, Honolulu, Hawaii, USA.

Davies, P. (2000). "Computerized Peer Assessment." Innovations in Education and Training International **37**(4): 346-355.

Davies, P. (2002). E-Assessment: Removing the C.A.A. boundaries. The 3rd Annual Conference of the LTSN Centre for Information & Computer Sciences, Loughborough University.

Davies, P. (2005). Weighting for Computerized peer-assessment to be accepted. 9th CAA International Computer Assisted Assessment Conference, Loughborough, UK.

Dierick, S. and F. Dochy (2001). "New lines in edumetrics: New forms in assessment lead to new assessment criteria." Studies in educational evaluation **27**: 307-329.

Elsen, E. R., Van den, H. J. A. Biemans, et al. (2005). Beoordeling in competentiegericht onderwijs: een praktische toepassing in het hoger onderwijs. In: Meten en Onderwijskundig Onderzoek. Onderwijs Research Dagen, Gent, Vakgroep Onderwijskunde Universiteit Gent.

Hamer, J., K. Kwong, et al. (2005). A method of automatic grade calibration in peer assessment. Seventh Australasian Computer Science Education Conference (ACE'2005), Newcastle,Australia.

Hughes, I. (2001). "But isn't this what you're paid for? The pros and cons of peer- and self-assessment." Planet Magazine(2): 20-23.

Keatley, C. and D. Kennedy (2004). Peer and Self Assessment, The National Capital Language Resource Center.

Langan, A. M. and C. P. Wheater (2003). "Can students assess students effectively? Some insights into peer-assessment." Learning and Teaching in action **2**(1).

Magin, D. and P. Helmore (2001). "Peer and teacher assessments of oral presentations: how reliable are they?" Studies in Higher Education **26**: 287-298.

Parsons, R., R. Handy, et al. (2003). Development of authoring software for peer and self assessment of text based exercises using the web as an interface. Teaching Development Fund, LTSN Centre for Bioscience, School of Biomedical Sciences, University of Leeds.

Prins, F. J., D. M. A. Sluijsmans, et al. (2005). "Formative peer assessment in a CSCL environment: a case study." Assessment & Evaluation in Higher Education **30**(4): 417-444.

Race, P. (1999). 2000 Tips for lecturers. London, Kogan Page.

Stefani, A. J., Ed. (1994). Self, peer and group assessment procedures. An enterprising curriculum: Teaching innovations in Higher Education. Belfast, HMSO.

Topping, K. (1998). "Peer assessment between students in colleges and universities." Review of Educational Research **68**(249-276).

Volder, M. d. (2005). ESPACE: Elektronisch Systeem voor Peer Assessment en Coaching Efficiency, Digitale Universiteit NL.