Can Online Peer Assessment be Trusted?

L'hadi Bouzidi[1] and Alain Jaillet[2]
[1]Department of Mathematics, Exact Science Faculty, A. Mira University, Bejaia, Algeria // lhadi_bouzidi@yahoo.fr
[2]ULP Multimedia, Strasbourg, France // alain.jaillet@ulpmm.u-strasbg.fr

ABSTRACT

The excessive workload generated by the assessment of exam papers in large classes and the need to give feedback in time often constitute a rather heavy burden for teachers. The online peer assessment can contribute to reduce this workload and, possibly, to improve learning quality by assigning the assessment task to students. However, this raises the question of validity. In order to study this question, we carried out an experiment of online peer assessment in which 242 students, enrolled in 3 different courses, took part. The results show that peer assessment is equivalent to the assessment carried out by the professor in the case of exams requesting simple calculations, some mathematical reasoning, short algorithms, and short texts referring to the exact science field (computer science and electrical engineering).

Keywords

Peer Assessment, Self-Assessment, Peer Assessment Validity, Item Analysis

Introduction

Learning assessment often constitutes a problematic task for the teacher. Its repetitive nature makes the teacher's job fastidious and the grading of hundreds of exam papers is far from exciting. In addition, the educational value of the assessment methods is often very low. The post-exam phases are difficult to put into practice with a large number of students, particularly, the feedback phase. Moreover, students are not really fond of these phases. In order to improve this situation, several devices can be used depending on the aspects that need to be changed. In this study, our primary concern is not to improve the reliability or the quality of assessment (De Ketele & Gerard, 2005), but rather to reduce the teacher's workload, without affecting the quality of assessment. In order to attain this, we suggest the use of computer-assisted assessments (Davies, 2001; Stephens & Bull, 1998) and the application of online peer assessment (Topping, 1998). Our proposal could be acceptable for the formative assessment of the learning process (Perrenoud, 2001) as this assessment method does not have many direct consequences on decisions concerning access to the following year or to certification. However, what about summative evaluation? Can grading carried out by students be trusted? What are the conditions of acceptability of our proposal? Several attempts (Topping, 1998) have been made to involve more students in this kind of approach, but it seems that the use of technologies can facilitate the implementation of these solutions. What remains to be established is whether this study can show that the assessment, thus produced, will be acceptable. Our research focuses on this issue of acceptability by measuring precisely the difference between assessments carried out by peers and those carried out by the teacher.

Online peer assessment

The use of online peer assessment is recent. It allows the students to assess the work of their peers and even assess their own work. Unlike the self-assessment techniques, which are generally limited to basic cognitive levels (Bloom, 1956; Anderson & Krathwohl, 2001), peer assessment enables to develop learning at high cognitive levels. This is a method that involves the students in the revision, assessment, and feedback process of work online. Doiron (2003) indicates that certain authors criticize the use of information and communication technology for peer assessment, arguing that it is not as rigorous as traditional types of assessment, that it requires too much student effort by putting too much pressure on them, that it is not reliable, and that it is not necessarily fair. In response, other authors such as Bostock (2000) are convinced that the (formative or summative) assessment of other students' work by the students themselves has several advantages for the learning process, both for the assessor and the assessee. Bostock points out that peer assessment encourages the students to be independent and develops skills in high cognitive areas. He acknowledges certain weaknesses of this type of assessment, in particular the over-estimating of friends' work, but he explains that this can be avoided by setting up a system which would guarantee anonymity, multiple assessments, a great number of assessors, and moderation by the teacher. Moreover, Bostock, specifies that Internet and information and communication technologies enable an easier management of a greater number of students.

Several authors (Rada & Michailidis & Wang, 1994; Fisher, 1999) explain that peer assessment may be proposed as an alternative solution to reducing teacher workload. Thus, Cho et al. (2006) explain that, despite the progress made during the last two decades in student writing skills, the courses rarely include real comprehensive writing tasks. According to them, this is due to the teachers' workload: assessing writing skills requires too much time and effort (Rada et al., 1994). They suggest resorting to peers in order to assess the students' work rather than systematically resorting to evaluation by teachers (Rada et al., 1994).

As regards the reliability of peer assessment, few studies have been carried out (Haaga, 1993; Falchikov, 1986; Falchikov & Goldfinch, 2000; Mowl & Pain, 1995; Stefani, 1994; Cheng & Warren, 1999). The majority of these studies have analyzed the performance in only one class (from 45 to 63 students) (Cho et al., 2006). It seems that only one study examined several courses with a significant number of participants (708) (Cho et al., 2006). Various measurements were used: comparison of means, Pearson's product moment correlation, concordance percentage, intra-class correlation, etc. The results are contradictory. There are good reasons for both confirming and contesting the reliability of marks awarded by peers. Hereafter, we will reveal the results of some of these studies.

Sadler & Good (2006) carried out a peer assessment experiment on 4 secondary school classes. By comparing the marks awarded by teachers with those awarded by pupils and with those obtained by self-assessment, they claim to have obtained a very high correlations (r=0.91 to 0.94). They also noticed that pupils slightly underestimate work of their comrades, whereas they overestimate their own. Regarding the impact of peer assessment on the learning process, they claim that self-assessment reinforces pupils' learning whereas peer assessment does not. Lastly, they explain that peer assessment considerably reduces the teachers' workload.

Zevenbergen (2001) believes that peer assessment is an efficient learning tool which can help mathematics teachers, but that should not be seen as an alternative to teacher's assessment of pupils' work.

Cho et al. (2006) carried out a peer assessment experiment on 16 classes (708 students) studying the science of education. The results reveal that peer assessment, based on at least 4 assessors per paper, is extremely reliable and is as valid as the teacher's assessment.

In summary, the literature review reveals that there are good theoretical reasons both for and against the reliability and validity of peer assessment (Cho et al., 2006). The previous empirical work is not large. Although the literature concludes that peer assessment appears valid (Falchikov & Goldfinch, 2000; Topping, 1998), this validity still needs to be addressed in a larger scale study using a common metric, across many courses and levels of students.

Reliability and validity issues

In practice, the assessment can pose validity and reliability problems. This may concern the tests, the assessment tools and the assessors (Brennan, 2001). The validity of a test is regarded as acceptable if the test comprises only items enabling the assessment of competences corresponding to training objectives. The reliability of an assessment tool is acceptable if it enables the accurate observation of the learning competence and makes it possible to give an opinion on this competence. An assessor is regarded as reliable if he or she has a good knowledge of the subject concerned by the assessment and if he or she also shows a good aptitude to making responsible judgments.

Validity and reliability remain debatable issues. Cho et al. (2006) note that validity is sometimes misreported in literature as reliability. This leads us to clarify these two concepts based on the definitions provided by Cho et al. (2006) and Falchikov et al. (2000). These authors regard the reliability of peer assessment as a variable that can be measured by the similarity between the marks given by peers. They consider the validity of peer assessment as a variable that can be measured by the similarity between the marks attributed by peers and by teachers.

Assumptions and objectives

We are considering tests and assessment tools that should guarantee validity and reliability. On the one hand, teachers are deemed pertinent assessors and are therefore considered reliable. On the other hand, we know that students are neither reliable nor valid assessors. Consequently, we wanted to find out whether several students taken together could be deemed as a valid "collective" assessor. In order to verify this, we compared the marks awarded by

the teacher and those awarded by the groups of students (the peers). If, on a significant sample and on the basis of a 95% confidence limit, we obtained that the marks awarded by the peers were the same as those awarded by the teacher, we would be able to confirm that the students' assessments could be trusted (i.e. that peer assessment is valid). This study is based on tests with closed or half-open questions (short calculations, mathematical reasoning, writing short algorithms, and drafting short texts).

Methodology

The validity of peer assessment can be verified by comparing peer assessment with the assessment carried out by the teacher (Cho et al., 2006). This assumes that:

- The teacher is a reliable assessor,
- The test items are set up correctly,
- The test constitutes a reliable assessment tool (internal consistency of items),
- The observation and marking tools (computer-assisted assessment tools in our case) are reliable,
- And the marking instructions and scale schemes are clear and coherent.

This has led us to use a method which requires that five hypotheses must be confirmed before checking the validity of peer assessment (see figure 1).

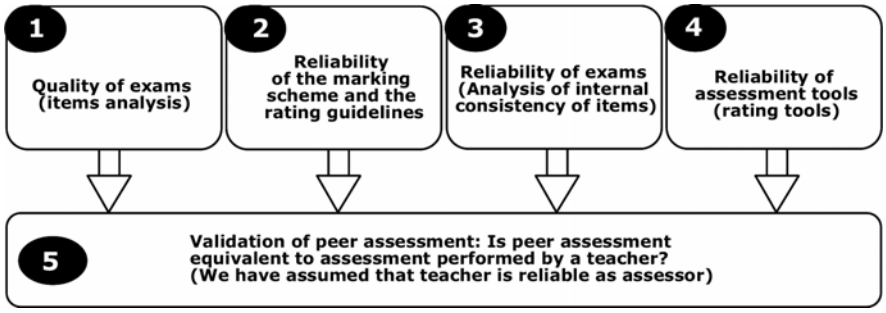

Figure 1. Validation process of peer assessment

As educational experts, teachers are considered reliable assessors. Proving that peer assessment is equivalent to teacher assessment could validate peer assessment. The validity of the tests can be checked through a comparative analysis of the papers of the same test by several experts (at least three) in the subject area (De Ketele et al., 2005). An item analysis (Wright & Panchapakesan, 1969) can also be used for validating the test. The reliability of the test can be checked by analyzing the internal consistency of its various items (Cronbach, 1951). In the case of computer-assisted assessments, it is more difficult to verify the reliability of the observation and marking tools. This difficulty is more important in the case of peer assessment. Human, technical and technological factors often contribute positively or negatively to this reliability. The marking instructions and scale schemes define the rules for awarding marks. Verifying their validity can be carried out by checking the coherence of the scale schemes (independent questions, accuracy) (Dubus, 2006) and by analyzing the correspondence between the question formulation and the marking instructions.

Constructing quality of test items: This was performed by an item analysis (Nunnally, 1967) although we did not carry out a complete quality analysis of the tests. We used this technique to identify some abnormalities in the test items. The level of these abnormalities enabled us to determine whether the test could be used to check the validity of peer assessment or not. We considered two metrics widely used in docimology: the difficulty index and the discrimination index (Girard, 2001). The comparison of the values of these statistical indexes resulting from marks awarded by peers with those resulting from marks awarded by the teacher could only explain certain abnormalities that will be taken into consideration in our decision regarding the validity of peer assessment.

The difficulty index "*p*" is the percentage of failure to answer the question. It is equal to the number of students having replied incorrectly, divided by the total number of students. When *p* is close to 0: this shows an easy question. When it is close to 1: this shows a hard question.

The discrimination index "*r*" (which varies from -1 to +1) represents the correlation (Pearson) between the answers to the question and the total for the other questions. Discrimination is considered (Girard, 2001) to begin at r=0.20. The items of average difficulty (p from 0.40 to 0.60) maximize discrimination. A discrimination coefficient equal to zero indicates that there is no discrimination. When it is negative, this shows incoherence: the best students fail the question, the weakest ones answer correctly. The Office of Educational Assessment of the University of Washington (Scorepak, 2005) classifies the discrimination coefficient as follows: $r > 0.30$ indicates that the question is good; r between 0.10 and 0.30 indicates that the question is acceptable; $r < 0.10$ indicates that the question is bad.

The 4 kinds of item abnormalities that have been used to check the test validity are exposed in Table 1, presented hereafter.

Table 1. Various situations of abnormal items

"*p*" (difficulty index)	"*r*" (discrimination index)	Abnormality	Observation
Between 0.40 and 0.60	Lower than 0.10	A1: Average difficulty and no discrimination	Generally speaking, questions of average difficulty maximize the discrimination of the students
. *p*	Between -0.10 and 0.10	A2: High negative discrimination	Negative discrimination means that the better students fail the question and the weaker ones answer correctly.
Higher than 0.80	Higher than 0.40	A3: Difficult question and good discrimination	Questions that are too easy or too hard should not discriminate the students
Lower than 0.20	Higher than 0.40	A4: Easy question and good discrimination	

Reliability of the test (Internal consistency of the items): This is the coefficient α of Cronbach (Cronbach, 1951; Bouchet, Guillemin, Hoang, Cornette & Briançon, 1996), which is the metric usually used. It is a statistical index that varies between 0 and 1 and enables the assessment of an assessment instrument's homogeneity (the internal consistency or coherence), made up of a series of items that should all contribute to understanding the level of knowledge or skill on a given theme. This index conveys a degree of homogeneity that increases when its value approaches 1. There is no known statistical distribution that makes it possible to conclude whether Cronbach's alpha is acceptable or not (McKell, 2000). The empirical limits resulting from the psychometrics used as a reference: In an exploratory study, Cronbach's alpha is considered as acceptable if it is between 0.6 and 0.8 (Zimmerman, Beverly, Sudweeks, Richard, Shelley, Monte et al., 1990; Evrard, Desmet, Dussaix, Lilien, Pras & Roux, 2003; Nunnally, 1967). The Office of Educational Assessment of the University of Washington (Scorepak, 2005) classifies the reliability coefficient as follows:

- $\alpha > 0.90$: Excellent reliability; at the level of the best standardized tests.
- $0.80 < \alpha < 0.90$: Very good for a classroom test.
- $0.70 < \alpha < 0.80$: Good for a classroom test. There are probably only a few items that would require improvements.
- $0.60 < \alpha < 0.70$: Somewhat low. The test requires to be supplemented by other measurements (other tests) to determine the grades. There are probably some items that would require improvements.
- $0.50 < \alpha < 0.60$: Suggests a need for revising the test, unless it is quite short (ten or less items). The test definitely needs to be supplemented by other measures (e.g., more tests) for grading.
- $\alpha < 0.50$: Questionable reliability. This test should not count as the main contribution to the course grade and needs to be revised.

The reliability of the observation and marking tools: In our experiment, the "workshop" module of the Moodle platform was used. This module is stable and presents a user-friendly environment. However, the marking tool (scale scheme) had a weakness. For the scale scheme that we had defined, the marking tools provided by this workshop were in the shape of option buttons varying from "excellent" to "very poor". Each button corresponded to a mark that the assessor could choose. This method can lead to errors.

Marking instructions and marking schemes: The marking schemes can comprise additive boxes and subtractive boxes (Dubus, 2006) that may coexist within the same total. The condition so that the marking scheme remains coherent is that all the boxes remain independent from each other, i.e. a subtraction box that has used up its credit cannot start nibbling the points scored in another box. Thus, a true marking scheme does not allow any compensation between sections. The idea of goal analysis, which underlies the construction of marking scales, contradicts the idea of compensation, because a weakness in skill *A* cannot be compensated by a strength in skill *B*.

The analysis that we carried out, in this research, does not aim to improve or change the marking scheme, but rather to check whether or not they have had an impact on the quality of the assessment (objectivity). We define the quality of a marking scheme with two characteristics: first, its accuracy and, second, the complete independence of its boxes. The accuracy of the marking scheme refers to its ability to help giving opinions on elementary skills. We think that, in order to guarantee the best objectivity, it is important to supplement the marking scheme by marking instructions. The validity of the marking instructions is defined by the clarity and coherence of the wording used for the items (questions). Their analysis may explain abnormal differences between assessors.

In this study, we will resort to checking the quality of the marking instructions only if the items considered contain abnormal differences between assessors.

Validity of peer assessment: It can be defined by the measurement of the agreement between two judges: the teacher and the peers. For qualitative variables, we can use the Kappa nonparametric test of Cohen (Cohen, 1960). For ordinal variables, we can use the Spearman Rho Coefficient (Kruskal, 1958). For quantitative variables (as in this research), we can use Pearson's correlation coefficient (Cho et al., 2006; Haaga, 1993) or the t-test (Mowl & Pain, 1995; Cheng and Warren, 1999; Stefani, 1994).

Experiment

We chose to implement a digitalization process to put the exam papers carried out in class online. This choice guaranteed an environment faithful to the exams actually carried out in class. The process that we implemented can be described according to the following stages (see figure 2):

1. Organize an exam in class.
2. Collect and then digitalize the exam papers.
3. Process the digitalized papers to ensure that the students remain anonymous.
4. Place the digitalized papers on the virtual campus for each student.
5. Prepare the activity on the virtual campus by indicating the marking scheme, assessment criteria and various parameters for assessing the marks and the distribution of the papers to the students.
6. Open the assessments (for the teacher and for the peers).
7. Hand-over the marks and the feedback to the students and open the discussion.

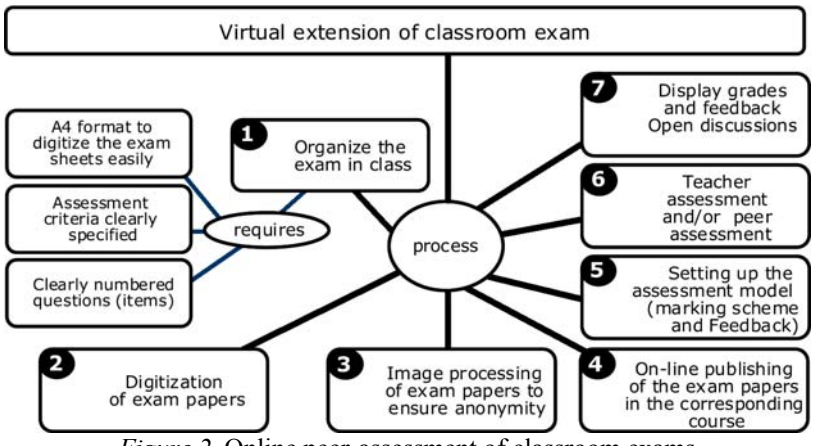

Figure 2. Online peer-assessment of classroom exams

In order to carry out our experiment, we based our research on five initial conditions:

- Subject area: exact sciences in order to restrict the scope of interpretation.
- Exam: performed in a supervised class in order to avoid any risk of invalidation.
- Marking scheme: clearly defined by the teacher.
- Test: including more than 9 questions.
- Average timing of the test: 2 hours.

The clear definition of the marking scales and the marking instructions were a condition on which we insisted. Moreover, we have recommended providing the exams solutions to the assessors.

In order to ensure that the students take this experiment seriously, we informed them that the quality of their assessment would be marked and that the assessment would be anonymous. Thus, they were invited not only to assess, but to correctly assess anonymous exam papers.

Data collection and treatment

There were 242 students (see table 2) of the second and the third year in engineering who took part in the experiment. Three exams were organized. They concerned three courses: computer architecture courses taught during the years 2006 and 2007 (CA-2006 and CA-2007), and the general electrical engineering course taught during the year 2007 (GEE-2007).

Table 2. Breakdown of students taking part in the experiment

Exams		Students		
Date	Identifier	Males	Females	Total
2006	CA-2006	36	32	68
2007	CA-2007	42	52	94
	GEE-2007	69	11	80
			Total:	242

The CA-2006 exam is composed of 17 items involving calculations (7 items) and drafting of short texts (one to five words: 5 items, one to five lines: 5 items). Six of these items enable the assessment of the basic cognitive levels: learning, understanding, and applying (Bloom, 1956; Anderson, 2001). The eleven other items enable the assessment of an additional cognitive level: analyzing.

The CA-2007 exam is composed of 13 items involving calculations (6 items), writing of algorithms (1 item) and drafting of short texts (one to five words: 2 items, one to five lines: 4 items). Three items enable the assessment of the low cognitive level: learning. Eight items enable the assessment of the basic cognitive levels. The two remaining items enable the assessment of an additional level: analyzing.

The GEE-2007 exam is composed of 9 items involving calculations and mathematical reasoning (establishment of a succession of mathematical transformations that lead to an expected formula). All the items enable the assessment of the four first cognitive levels (learning, understanding, applying, and analyzing).

All of the exams are marked out of 20 points. Their items are marked out of 0.5, 1, 1.5, 2, 2.5 and 3 points according to the marking scales. The smallest marks that can be awarded are 0.25 and 0.5 according to the exam.

For each test and for each paper, we gathered the following marks (see figure 3): one self-assessment mark, four marks awarded by peers, and one mark awarded by the teacher; this applies for each item. Once this data was gathered, we added up the marks linked to the items (questions) in order to calculate the marks for all of the assessors:

- “TM”: Marks awarded by the teacher.
- “SM”: Marks produced by self-assessment.
- “PM”: Marks awarded by peers disregarding the self-assessment.

“SPM”: Marks awarded by peers, incorporating the self-assessment.

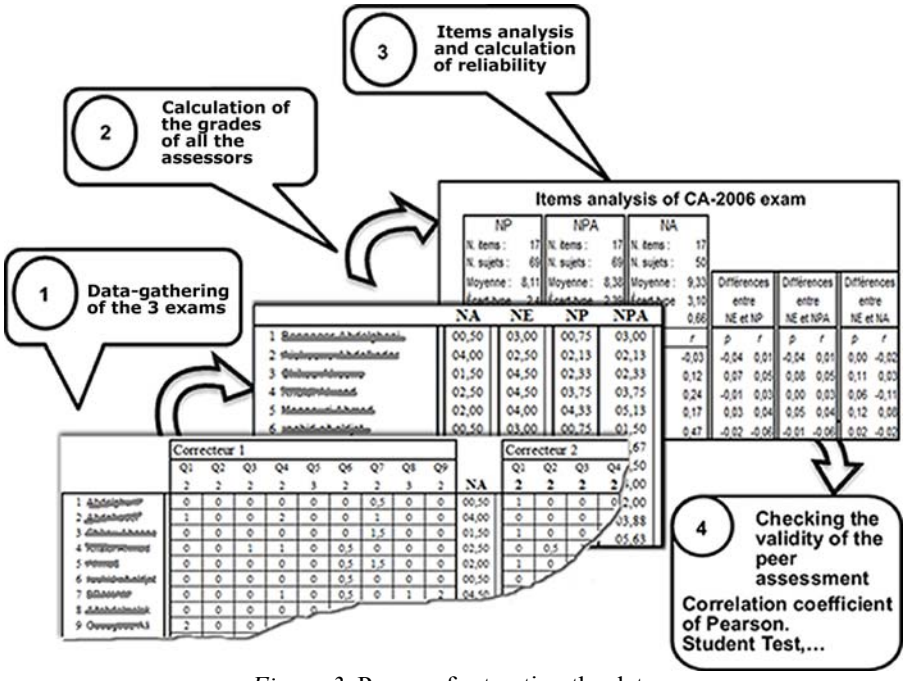

Figure 3. Process for treating the data

For each test and for each type of assessment (self-assessment, peer assessment, peer assessment incorporating self-assessment, and assessment by the teacher), we have performed an item analysis. In particular, we have calculated, for each type of assessment and for each test, the following indicators: averages and standard deviations of the marks and Cronbach's coefficient α. We have also calculated the difficulty index and the discrimination index for each item.

Analysis and interpretation

In relation to CA-2006 exam, the item analysis produced the results shown in Table 3. We notice an "A3 abnormality" (see Table 1) in Q16 and Q17 questions. However, for these questions, we observe a good agreement between the teacher's assessments and the peers' assessments. We also note that it is the peer assessment incorporating the self-assessment that is the closest to the assessment by the teacher. The students, with a positive average difference between their marks and those of the teacher (0.12), very slightly under-mark the papers of their peers. On the other hand, the authors of the papers, with a negative average difference between their marks and those of the teacher (-0.83), over-mark their papers.

The difficulty of the other questions, according to the teacher, is generally equal to that found by the peers (correlation equal to 0.97). However, there is an incoherence in the Q14 question that shows a difference of 26% between the teacher's severity and that of the peers (severity is considered here equivalent to the difficulty index). As this question appears to be valid, we think that this incoherence is not due to chance but rather to a poor definition of the marking instructions, which misled the students in their assessments. However, with an internal consistency of the marks (Cronbach's α) equal to 0.61 (teacher's view), we can consider this test as relevant for a peer assessment validity analysis.

Concerning the CA-2007 exam, the item analysis generated the results shown in Table 4. This table shows that the students marked the answers to question Q7 severely (p=0.74), whereas the teacher marked the answers to the same question with average severity (p=0.59). The difference is too large to be explained by chance alone. Another explanation is a poor indication of the marking instructions for this item. Indeed, in this question, the second and third parts of the answer expected are based on the result of the first part. This causes a problem with regard to the independence of the different parts of the marking scheme. The teacher's grading takes into account the second and

third ability even if the first one is not right. Since the marking instructions given to the students do not specify this, the students applied incoherent marking rules, different from those of the teacher. This automatically distorted our experiment, which was based on everyone using the same marking scale and marking rules. However, for the other questions in the test, we can consider that our experiment respected the hypothesis of clear marking scale and marking instructions. We will therefore disregard question "Q7" while performing the peer assessment validity analysis.

Table 3. Item analysis of the CA-2006 exam

	TM Mean : 8,5 *SD* : 2,8 α : 0,61		*PM* Mean : 8,11 *SD* : 2,4 α : 0,55		*SPM* Mean : 8,38 *SD* : 2,39 α : 0,58		*SM* Mean: 9,33 *SD* : 3,10 α : 0,66		*NI* = 17 , *NS* = 69					
									Differences Between *TM* & *PM*		Differences Between *TM* & *SPM*		Differences between *TM* & *SM*	
Item	*p*	*r*	*p*	*r*	*P*	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*	*r*
Q1	0.12	-0.05	0.16	-0.06	0.15	-0.06	0.12	-0.03	-0.04	0.01	-0.04	0.01	0.00	-0.02
Q2	0.59	0.15	0.52	0.10	0.52	0.10	0.49	0.12	0.07	0.05	0.08	0.05	0.11	0.03
Q3	0.16	0.13	0.17	0.10	0.16	0.10	0.10	0.24	-0.01	0.03	0.00	0.03	0.06	-0.11
Q4	0.86	0.25	0.82	0.21	0.81	0.21	0.74	0.17	0.03	0.04	0.05	0.04	0.12	0.08
Q5	0.64	0.45	0.66	0.51	0.66	0.51	0.62	0.47	-0.02	-0.06	-0.01	-0.06	0.02	-0.02
Q6	0.71	0.22	0.68	0.12	0.68	0.12	0.74	0.19	0.03	0.10	0.03	0.10	-0.03	0.03
Q7	0.84	0.22	0.81	0.28	0.80	0.28	0.80	0.21	0.03	-0.06	0.04	-0.06	0.04	0.01
Q8	0.88	0.02	0.80	0.07	0.77	0.07	0.68	0.17	0.08	-0.05	0.11	-0.05	0.20	-0.15
Q9	0.71	0.39	0.76	0.32	0.75	0.32	0.72	0.37	-0.05	0.07	-0.04	0.07	-0.01	0.02
Q10	0.67	-0.06	0.70	0.03	0.68	0.03	0.65	0.10	-0.03	-0.09	-0.02	-0.09	0.02	-0.16
Q11	0.27	0.23	0.34	0.23	0.33	0.23	0.26	0.25	-0.07	0.00	-0.05	0.00	0.02	-0.02
Q12	0.37	0.12	0.40	0.15	0.38	0.15	0.28	0.26	-0.03	-0.03	0.00	-0.03	0.09	-0.14
Q13	0.59	0.50	0.60	0.52	0.58	0.52	0.58	0.66	-0.01	-0.02	0.01	-0.02	0.01	-0.16
Q14	0.16	0.25	0.42	0.21	0.39	0.21	0.28	0.30	-0.26	0.04	-0.24	0.04	-0.12	-0.05
Q15	0.78	0.20	0.79	0.27	0.78	0.27	0.72	0.25	-0.01	-0.07	0.01	-0.07	0.06	-0.05
Q16	0.88	0.55	0.92	0.39	0.91	0.39	0.84	0.29	-0.04	0.16	-0.02	0.16	0.04	0.26
Q17	0.93	0.40	0.89	0.32	0.88	0.32	0.82	0.45	0.03	0.08	0.04	0.08	0.11	-0.05
μ	0.60	0.23	0.61	0.22	0.60	0.22	0.55	0.26	-0.02		0.00		0.04	
Correlation between severities (difficulty indexes)									0.97		0.97		0.97	
Means differences									0.39		0.12		-0.83	

Legend		
	p: Difficulty index *r*: Discrimination index *α*: Cronbach's coefficient (reliability) μ: Means *TM*: Teacher marks *PM*: Peer marks	*SPM*: Marks awarded by Peers and by self-assessment *SM*: Marks awarded by self-assessment *NI*: Number of items *NS*: Number of exam sheets *SD*: Standard deviation

Table 4. Item analysis of the CA-2007 exam

	TM *NS*=94 Mean: 11.03 *SD*: 3,00 α: 0,65		*PM* *NS*=94 Mean: 10.50 *SD*: 2,60 α: 0,67		*SPM* *NS*=94 Mean: 10.80 *SD*: 2,63 α: 0,68		*SM* *NS*=89 Mean: 12,00 *SD*: 2,90 α: 0,58		*NI* = 13					
									Differences Between *TM* & *PM*		Differences between *TM* & *SPM*		Differences between *TM* & *SM*	
Item	*p*	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*	*r*
Q1	0.24	0.29	0.26	0.21	0.25	0.23	0.20	0.21	-0.02	0.08	-0.01	0.06	0.04	0.08
Q2	0.04	0.13	0.06	0.16	0.05	0.14	0.03	0.00	-0.02	-0.04	-0.01	-0.02	0.00	0.13
Q3	0.20	0.30	0.24	0.38	0.24	0.38	0.23	0.25	-0.04	-0.08	-0.04	-0.09	-0.03	0.04
Q4	0.39	0.36	0.38	0.40	0.38	0.40	0.34	0.29	0.01	-0.04	0.01	-0.04	0.04	0.07
Q5	0.64	0.44	0.66	0.38	0.64	0.40	0.56	0.34	-0.01	0.05	0.00	0.04	0.08	0.09
Q6	0.48	0.16	0.48	0.28	0.47	0.28	0.45	0.17	0.00	-0.13	0.00	-0.12	0.03	-0.01
Q7	0.59	0.24	0.74	0.34	0.72	0.34	0.63	0.21	-0.15	-0.09	-0.13	-0.10	-0.04	0.04

Q8	0.14	0.03	0.13	0.09	0.14	0.07	0.13	-0.04	0.00	-0.06	0.00	-0.04	0.00	0.07
Q9	0.59	0.34	0.59	0.32	0.58	0.31	0.54	0.27	0.00	0.02	0.01	0.03	0.05	0.08
Q10	0.40	0.50	0.39	0.49	0.35	0.53	0.24	0.32	0.01	0.00	0.05	-0.03	0.16	0.17
Q11	0.78	0.27	0.84	0.22	0.82	0.25	0.77	0.29	-0.06	0.04	-0.05	0.02	0.01	-0.02
Q12	0.57	0.23	0.53	0.35	0.52	0.37	0.47	0.44	0.04	-0.12	0.05	-0.14	0.10	-0.21
Q13	0.53	0.50	0.53	0.45	0.50	0.49	0.40	0.49	0.00	0.05	0.03	0.01	0.13	0.01
μ	0.43	0.29	0.45	0.31	0.44	0 .32	0.39	0.25	-0.02		-0.01		0.04	
Correlation between severities (difficulty indexes)									0.98		0.98		0.96	
Means differences									0.54		0.23		-0.96	

Legend	*p*: Difficulty index *r*: Discrimination index α: Cronbach's coefficient (reliability) μ: Means *TM*: Teacher marks *PM*: Peer marks	*SPM*: Marks awarded by Peers and by self-assessment *SM*: Marks awarded by self-assessment *NI*: Number of items *NS*: Number of exam sheets *SD*: Standard deviation

As regards the GEE-2007 exam, the item analysis showed low reliability (0.41). Consequently, we could not use it to verify the peer assessment validity.

Checking the validity of peer assessment: Although the item analysis has already given us an idea of this validity, we preferred reinforcing our analysis with other metrics already used in other research dealing with the same issues as ours (Cho et al., 2006; Haaga, 1993; Mowl & Pain, 1995; Cheng & Warren, 1999; Stefani, 1994). These metrics (see Table 5) are Pearson's correlation coefficient and the "T-Test". We used these two metrics in the comparison between the teacher's marks (TM) and those awarded by the peers (PM: peer marks disregarding self-assessment, SPM: peer marks incorporating the self-assessment, and SM: self-assessment mark). It should be noted that we did not take into account the GEE-2007 exam and the question Q7 of the CA-2007 exam, which were rejected by the item analysis.

The correlation coefficients between the teacher's marks and those of the peers are high (0.90±2). The "T-Test", applied to the marks awarded by the teacher and the marks awarded by peers, is positive. We can confirm that peer assessment is valid for the category of exams targeted in our research and that the assumptions outlined above are verified.

Table 5. T-test and correlation coefficient between the teacher marks and the peer marks.

	CA-2006 exam		CA-2007 exam without Q7	
	TM & *PM* Comparison	*TM* & *SPM* Comparison	*TM* & *PM* Comparison	*TM* & *SPM* Comparison
Number of papers	68	68	94	94
Variable 1	*TM*	*TM*	*TM*	*TM*
Mean	8.426	8.426	9.806	9,806
Standard deviation	2.830	2.830	2.651	2,651
Variable 2	*PM*	*SPM*	*PM*	*SPM*
Mean	8.1265	8.4009	9.7258	9,9566
Standard deviation	2.3686	2.4082	2.3505	2,3331
T-Test Hypotheses: *H0*	$\mu_{TM} = \mu_{PM}$	$\mu_{TM} = \mu_{SPM}$	$\mu_{TM} = \mu_{PM}$	$\mu_{TM} = \mu_{SPM}$
Signification level (%)	5%	5%	5%	5%
Number of degrees of freedom	67	67	93	93
T-Score found	1.8529	0.1886	0.6404	1.3470
Critical value	1.9944	1.9944	1.9867	1.9867
Conclusion	*H0* accepted	*H0* accepted	*H0* accepted	*H0* accepted
Correlation coefficient (Pearson)	0.88	0.91	0.89	0.91
d of Cohen	0.11	0.11	0.03	0.06
Effect size *r*	0.06	0.06	0.02	0.03
Legend :*TM*: Teacher Marks				

PM: Peer Marks
SPM: Marks award by peers and authors of exam papers (Self and Peer Marks)
μ_{TM}: Mean of the marks awarded by the teacher
μ_{PM}: Mean of the marks awarded by the peers
μ_{SPM}: Mean of the marks awarded by the peers and authors of exam papers

Our results compared to other research

Falchikov and Goldfinch (2000) performed a meta-analysis comparing the marks of peers to those of the teacher. This analysis does not deal with online peer assessment and mainly concerns undergraduate students. We have extracted some of the results from this meta-analysis by filtering the studies in relation to our research. These results prove that it is difficult to give an opinion on the validity of peer assessment while basing the research on previous studies. Indeed, the correlation coefficient between the peer marks and those of the teacher varies from 0.29 to 0.94. Peer assessment validity depends on several factors such as the reliability of the marking schemes and the clarity of the marking instructions. Our study proposes a method that makes it possible to avoid situations that could distort interpretation.

Falchikov and Goldfinch (2000) found that peer assessments resemble more closely teacher assessments when global judgments, based on well understood criteria, are used rather than when marking involves assessing several individual dimensions. Nevertheless, our study shows that it is possible, with an assessment involving several individual dimensions, to obtain a very good correlation between the peer marks and those of the teacher. These results, however, are limited to exams that mainly contain questions referring to exact science field and requesting simple calculations, some mathematical reasoning, short algorithms, and short texts.

Moreover, online peer assessment has been used during the past decade and some studies have dealt with the issue of validity (Cho et al., 2006; Tseng & Tsai, 2007). Our study confirms the results obtained by these studies. Indeed, Tseng et al carried out a study on the validity of peer assessment in a computer course involving 184 students. They claimed that the marks awarded by the peers largely corresponded to those awarded by the experts, thus indicating that peer assessment in high schools could be considered as a valid method of assessment. In addition, Cho et al (2006), while carrying out an experiment involving 708 students, proved that the overall assessments of at least 4 peers on a written examination are reliable and just as valid as the assessments made by the teacher. The results suggested that the teachers' concerns regarding the reliability and validity of peer assessments should not prevent a peer assessment from being applied, at least with the appropriate scaffolding.

Conclusion

This study suggests trusting peer assessment when it is applied to exams that contain questions referring to the exact science field (calculations, mathematical reasoning, short algorithms and drafting of short texts) and when the exam paper is marked by at least four peers. The results also show that a combination of peer assessment with self-assessment gives a better validity to this assessment method. However, one must take some precautions (the quality of the marking schemes, the clarity of questions, the grading of the assessments quality, the technical assistance, and the assessment of examples).

The choice of using exams already carried out (in a traditional manner) in the classroom makes it possible to attain an authentic assessment and opens educational and administrative horizons (statistical treatment, item analysis, validity analysis, etc.). In particular, this method helps teachers identify what is wrong in their exams, thus giving them an opportunity to correct them. However, with respect to the teacher's workload, this choice is a disadvantage. Consequently, we must supplement our solution by creating a system that enhances online assessment by an automatic phase that makes it possible to put exams carried out in classroom online (this system is being developed).

Concerning the educational aspect, our research will be supplemented by an analysis of the interaction occurring between all of the actors (students and teachers) participating in the experiment. This will give us information on the impact that this assessment method has on the learning process.

Finally, regarding the scope of this research, our work will be supplemented by an experiment involving more students and courses. We will also explore the relationship that exists between peer assessment validity and variables like the number of assessors, the scale schemes, the student knowledge level, the knowledge fields to which the exams belong, and the cognitive levels required by the exam items.

References

Anderson, L., & Krathwohl, D. (2001). *A Taxonomy for learning teaching and assessing: A revision of Bloom's taxonomy of educational objectives*, New York: Wesley Longman.

Bloom, B. (1956). *Taxonomy of Educational objectives: The Classification of Educational Goals*, New York: Longmans Green.

Bostock, S. (2000). *Student Peer Assessment, Learning Technology*, retrieved July 29, 2009, from http://www.keele.ac.uk/depts/aa/landt/lt/docs/bostock_peer_assessment.htm.

Bouchet, C., Guillemin, F., Hoang Thi, T., Cornette, A., & Briancon, S. (1996). Validation du questionnaire St Georges pour mesurer la qualité de vie chez les insuffisants respiratoires chroniques. *Revue des maladies respiratoires, 13* (1), 43-46.

Brennan, R. (2001). An Essay on the History and Future of Reliability from the Perspective of Replications. *Journal of Educational Measurement, 38* (4), 295-317.

Cheng, W., & Warren , M. (1999). Peer and Teacher Assessment of the Oral and Written Tasks of a Group Project. *Assessment & Evaluation in Higher Education, 24* (3), 301-314.

Cho, K., Schunn, C., & Wilson, R. (2006). Validity and Reliability of Scaffolded Peer Assessment of Writing From Instructor and Student Perspectives. *Journal of Educational Psychology, 98* (4), 891-901.

Cohen , J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement, 20* (1), 37-46.

Cronbach, L. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16* (3), 297-334.

Davies, P. (2001). *Computer Aided Assessment MUST be more than multiple-choice tests for it to be academically credible?* retrieved July 29, 2009, from http://www.caaconference.com/pastConferences/2001/proceedings/index.asp.

De Ketele, J., & Gerard, F. (2005). La validation des épreuves d'évaluation selon l'approche par les compétences. *Mesure et évaluation en éducation, 28* (3), 1-26.

Doiron, G. (2003). The Value of Online Student Peer Review, Evaluation and Feedback in Higher Education. *CDTL Brief, 6* (9), 1-2.

Dubus, A. (2006). *La notation des élèves – Comment utiliser la docimologie pour une évaluation raisonnée*, Paris: Armand Colin.

Evrard, Y., Pras, B., Roux, E., Desmet, P., Dussaix, A., & Lilien, G. (2003). *Market. Etudes et recherches en marketing*, Paris: Dunod.

Falchikov, N. (1986). Product comparisons and process benefits of collaborative self and peer group assessments. *Assessment & Evaluation in Higher Education, 11* (2), 146–166.

Falchikov, N., & Goldfinch, J. (2000). Student Peer Assessment in Higher Education: A Meta-Analysis Comparing Peer and Teacher Marks. *Review of Educational Research, 70* (3), 287-322.

Girard, M., & Normand, S. (2001). Guide de lecture d'un rapport d'analyse d'items, retrieved July 29, 2009, from http://medecine.ulb.ac.be/tools/docimo/Guide%20de%20lecture%20AnItem.pdf.

Haaga, D. (1993). Peer review of term papers in graduate psychology courses. *Teaching of Psychology, 20* (1), 28–32.

Kruskal, W. (1958). Ordinal measures of association. *Journal of the American Statistical Association, 53*, 814–861.

McKell, C. (2000). Mise en œuvre et évaluation de certaines mesures énoncées dans le Plan des services de réadaptation du Nouveau-Brunswick (NB102), retrieved July 29, 2009, from http://gnb.ca/0383/HTF/index-f.asp.

Mowl, G., & Pain, R. (1995). Using self and peer assessment to improve students' essay writing - A case study from geography. *Innovations in Education and Training International, 32* (4), 324–335.

Nunnally, J. (1967). *Psychometric theory*, New York: McGraw Hill.

Perrenoud, P. (2001). Évaluation formative et évaluation certificative, des postures définitivement contradictoires ? *Formation professionnelle suisse, 4*, 25-26.

Rada, R., Michailidis, A., & Wang, W. (1994). Collaborative hypermedia in a classroom setting. *Journal of Educational Multimedia and Hypermedia, 3* (1), 21–36.

Sadler, P., & Good, E. (2006). The impact of self- and peer-grading on student learning. *Educational Assessment, 11* (1), 1-31.

Scorepak (2005). Understanding Item Analysis Reports, retrieved July 30, 2009, from http://www.washington.edu/oea/services/scanning_scoring/scoring/item_analysis.html.

Stefani, L. (1994). Peer, self and tutor assessment: Relative reliabilities. *Studies in Higher Education, 19* (1), 69–75.

Stephens, D., Bull, J., & Wade, W. (1998). Computer-assisted assessment: suggested guidelines for an institutional strategy. *Assessment and Evaluation in Higher Education, 23* (3), 283–294.

Topping, K. (1998). Peer Assessment Between Students in Colleges and Universities. *Review of Educational Research, 68* (3), 249-276.

Tseng, S., & Tsai, C. (2007). On-line peer assessment and the role of the peer feedback: A study of high school computer course. *Computers & Education, 49* (4), 1161-1174.

Wright, B., & Panchapakesan, N. (1969). A Procedure for Sample-Free Item Analysis Wright and Panchapakesan. *Educational and Psychological Measurement, 29* (1), 23-48.

Zevenbergen, R. (2001). Peer assessment of student constructed posters: assessment alternatives in preservice mathematics education. *Journal of Mathematics Teacher Education, 4* (2), 95-113.

Zimmerman, B., Sudweeks, R., Shelley, M., & Wood, B. (1990). How to Prepare Better Tests: Guidelines for University Faculty, retrieved July 29, 2009, from http://testing.byu.edu/info/handbooks/bettertests.pdf.