Correlating peer and tutor-assessment on a low-stakes engineering assignment

Diogo Montalvão[1] and Thomas Baker

School of Engineering and Technology, University of Hertfordshire, United Kingdom

Abstract

Peer-assessment has been a subject of great debate in recent years. The way students perceive assessment and what motivates them when assessing, may differ significantly from the tutor. This paper discusses a study designed to correlate student's marking with the marks awarded by their tutors when peer-assessing one another from in-class oral presentations. A new and alternative approach to correlate results is presented, which is based on the normalisation of the quantitative judgements based on determined criteria. The methodology was *blind* and *holistic*, as described in previous works: some guidelines were provided to the students on what is considered acceptable without getting into detail (*holistic* marking) and peer-assessment marks were made confidential (*blind* approach). It was observed that students have a tendency to overrate

[1] **Corresponding author:** Diogo Montalvão, School of Engineering and Technology, University of Hertfordshire, College Lane Campus, Hatfield AL10 9AB, UK. Email: d.montalvao@herts.ac.uk

fellow students - especially where lower marks might be awarded. There is, however, direct agreement with the tutor's marking in terms of qualitative judgements, which is highlighted by the presented correlation method used to adjust students marks. The presented methodology to correlate marks between the students and tutor showed to be a promising one. After processing the data with this simple and straightforward algorithm, peer- and tutor assessment practically showed a perfect match.

Key words peer-assessment; assessment for learning; higher education assessment; assessment normalisation.

Introduction

Orsmond (2011), in his review on assessment within the higher education context, discriminates three ways of considering assessment in relationship to learning: assessment *of* learning, assessment *for* learning and assessment *as* learning. The widely accepted practice for assessment still is in terms of being *of* learning. Typically, students are assessed through summative assessment, e.g., an end of module examination in which the student's amount of learning is fully assessed by the tutor. However, Boud and Falchicov (2006) consider this approach as short-termed under the context of higher education, in which new material must be gradually appreciated through an iterative updating process, so that life-long self-learning is stimulated after graduation.

The conceptual shift of assessment from "*of* learning" to "*for* learning" and then to "*as* learning" can be traced back to Gipps (1994) and Earl (2003), respectively. Although with different emphasis, both acknowledge the introduction of student self-

assessment, in which the student assumes an important role in the assessment and judgment of his own work. Another frequently adopted method in which students are involved in the appreciation and appraisal of learning is peer-assessment (Van Gennip et al. 2010). According to Orsmond (2011), "peer-assessment shares most of the key features of self-assessment" and "feeds self-assessment activities particularly through the act of receiving and giving feedback." Prompt feedback is appreciated by students, because it gives them an early opportunity to assess their understanding of the course material, as acknowledged by Cotner et al. (2008) when introducing scratch-off immediate feedback assessment technique (IF-AT) forms and classroom response system (clickers) in large cohorts of students.

For peer-assessment to work effectively, some authors argue that one of the requirements is that the student and the tutor discuss and agree – or at least clarify and reflect on - assessment criteria (Falchikov and Boud 1989). Tsivitanidou et al. (2011), on a study investigating secondary school students' unmediated peer assessment skills, observed that students already have the basic skills required for the implementation of peer-assessment, although they recognised that "reciprocal peer assessment is a rather complex procedure that requires well developed peer assessment related skills" and that "students need to receive explicit training". However, Norton (2004) points out that students may become strategic learners who end up selecting what they identify as being the key aspects to obtain marks instead of spreading their attention to the whole span of the learning range. On the other hand, one might argue that students tend not to perceive assessment criteria as tutors do. Most often, they are not provided with the proper tools that would enable them to self-monitor their own learning (Hatzipanagos and Rochon 2010) and they still have little involvement in assessment design, marking and evaluation (Orsmond 2011).

Nevertheless, peer-assessment is an approach to learning that is nowadays recognised as a tool capable of getting the students more involved, making them take more responsibility, encouraging self-critical analysis and self-evaluation of student work and fostering debate and communication (Hunter 1999 and Dochy and McDowell 1997). Even if a poor correlation between students self-assessed marking and the tutor marking is observed, some authors (e.g., Kirby and Downs 2007) continue advocating in favour of self-assessment for learning – which is strongly correlated to peer-assessment according to Orsmond (2011).

Peer-assessment can be used as a tool to give students an opportunity to learn by making them aware of the different approaches employed by other students: an aspect highlighted by Van Gennip et al. (2010) by mentioning the social and interpersonal bonds involved in peer-assessment, from which students, by assessing one another, are actually learning from their peers. For example, Chuck and Young (2004) designed a formative assessment task to improve the scientific report writing skills of university students. This used a combination of peer- and self-assessment against specific criteria, where students were required to submit an amended report. It was shown that those that participated in this cohort-driven assessment task got better results on average compred with those that would have been obtained after the first submission.

With the above points in mind, a peer-assessment assignment was introduced in the modules of Technical Design and Geometrical Modelling I and II (1st year students) of Engineering undergraduate degrees taught in **institution removed for blind review**.

The peer-assessment case study discussed in this paper took place during the semester's last week (before examinations) and was focused on a part of the student's group assignment. The students were invited to mark one another's oral presentations

on sheets of paper (similar to voting bulletins). These were later returned to the tutor in such a way that the peer-assessment marks were kept anonymous.

Using the terminology from Russel et al. (2006), the followed methodology was *blind* and *holistic* as opposed to *open* and *category-based*. These authors describe these terms the following way: in a *blind* approach, students respond to the peer-assessment task in complete isolation from one another (anonymous marking), whereas in an *open* approach students discuss their thoughts and negotiate marks when peer-assessing within a group. In a *holistic* marking criteria style, students are asked to evaluate their peers' contribution and performance by looking at the 'big-picture'. Some guidelines can be provided to the students on what is considered acceptable, but the students are not asked to explicitly respond to them in any detail. On the other hand, on a *category-based* marking scheme, students are expected to mark their peers' performance against a specific set of marking criteria.

Results show that students are perfectly capable of making the correct distinctions between one another when speaking in qualitative terms. The correlation between the peer- and the tutor assessment in quantitative terms is not that different as well. However, a tendency to award generous marks to weaker students is observed.

An alternative approach to analyse the peer-assessment marks is presented as well, in an attempt to better understand how peer- and tutor assessment correlate in the quantitative judgments. This is based on the normalisation of the peer-assessment marks based on range and one parameter target (in this case, a value for the average was set as target). For the correlation to be comparable the average of the peer-assessment marks was set to the same value as the average of the marks from the tutor. The range was stretched in order to consider all the possible marking range as students showed to give marks on a narrower band than the tutor.

Method

Data sample

This study was carried out with data from **institution removed for blind review**, and the assessments were part of the "Technical Drawing and Geometrical Modelling I and II" modules. These are 1st year undergraduate modules taught in semesters A and B of the Engineering degrees. The data analysed corresponds to four cohorts of approximately 25 students each (a total of approx.. 100 students).

General assessment criteria

Assessment of these modules is based on a weighted combination of individual coursework (both in-class and homework) and a group design project that includes both an in-class presentation and a written report which is submitted at the end of the semester. This is illustrated in figure 1 pie chart, in which the lower half corresponds to group work and the upper half corresponds to individual work.

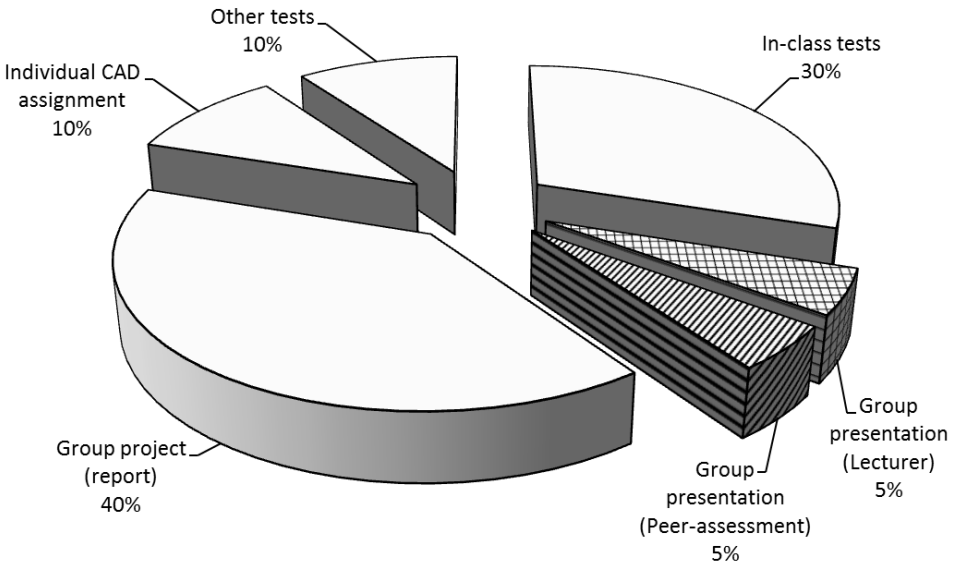

Figure 1. Marking distribution.

The group component of the assessment includes a report on the group project worth 40% of the final marks and a group presentation worth 10%. The group oral presentation is the average value between the marks given by the students when peer-assessing (worth 5%) and the marks given by the lecturer (worth another 5%). Although this was a group presentation, students were also assessed individually according to how well they performed on the roles they played as described below (each group was composed by 2~3 students, depending on the complexity of their projects).

Peer-assessment marking scheme

The in-class presentations ran during the semester's last week. Each group, composed of 2 to 3 students, prepared a presentation using both the CAD software and a slideshow. Each student was given up to 5 minutes to make their presentation regarding their contribution to the group work.

Both the teacher and the students based their assessment on the same marking scale, ranging from 1 to 10 (figure 2): three levels of pass and two levels of fail were considered. The zero marks were reserved for students who did not attend the presentations (any student presenting would get at least a minimum of 1 mark).

Figure 2. Marking range for the peer-assessment.

No specific marking criteria were given to the students other than to assess the specific objectives defined in the coursework briefing. To elaborate a list of criteria

would make the task unfeasible considering time constraints and the number of oral presentations each student had to assess. There also are other factors that were taken into account on this decision. First, the students might feel it is the job of the lecturer (and not students) to mark work, especially if a considerable amount of time is needed, as pointed earlier by Falchikov (2003). Secondly, according to a study from Brown et al. (1998), students tend to struggle when assessing via a discriminated marking scheme, even if they already have a general perception of the quality of the work. Thus, an *holistic* approach seemed more appropriate to the oral presentations, in which students look at the 'big-picture'. Furthermore, this was a low-stakes part of the assignment aiming at giving some guidance to the students on the progress of their work from the tutor's comments.

For the peer-assessment to be as *unbiased* as possible, some precautions were taken:

1. To ensure that students take it seriously and engage with the exercise, and to avoid cases where students might show favoritism (e.g. by marking certain students with all 10's (or 1's)), they were told that any "non-differentiating" marking sheets would be rejected.
2. It was guaranteed that each individual marking sheet would be kept confidential and would be destroyed after the results were collected. – *blind* methodology;
3. To make the students feel this was their free choice and not an imposition, they were told they could refrain from assessing their peers and submit the marking sheet blank;
4. To provide a sense of justice and fairness, the students could opt out from being assessed by their peers. In such a case, the lecturer's assessment would be worth the whole 10% of the presentation marks.

From the almost 100 students participating in this study, none fitted the 3rd and 4th categories listed above. However, there were a few whose marking sheets were rejected for the reasons described in point 1 above. With respect to the 2nd point, it is not possible to tell if the students feared discrimination because those sitting next to them could possibly peek into their marking sheets.

Results and discussion

General results

A histogram plot displaying the number of times (frequency) a grade was given by both the lecturer and the peers[2], is presented on figure 1. This picture includes the results for both the modules together showing a trend that follows what resembles a normal (Gaussian) distribution. The decision to join both the modules in a single plot is justified from the fact that the conclusions that might be drawn from each of the modules alone would be essentially the same. Furthermore, the statistical significance is improved, since the sample so obtained increased to a size of N=93 elements.

[2] Marks were rounded to an integer value. For instance, grade 7 in the histogram bar chart includes all marks ranging from 6.5 to 7.4.

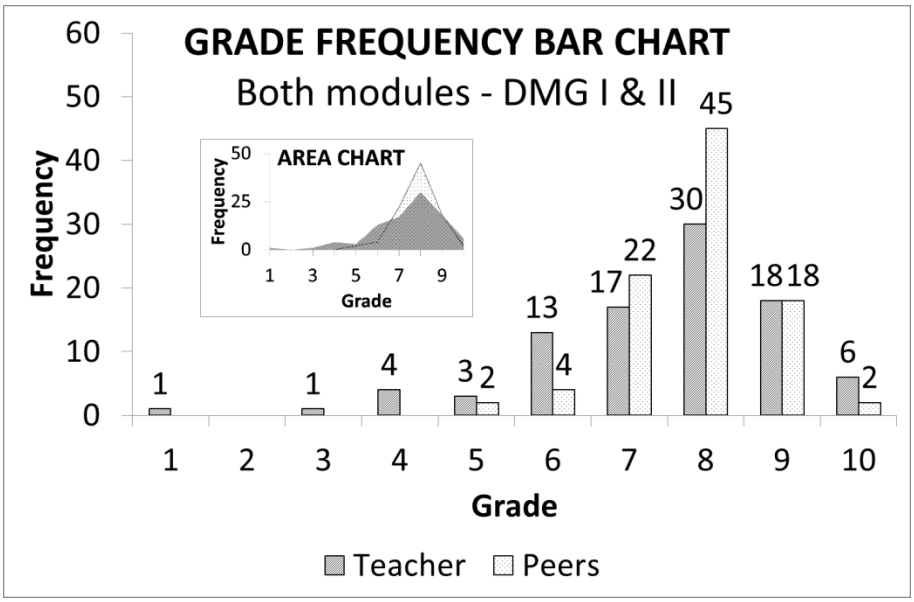

Figure 1. Histogram showing the amount of times a grade was given.

A smaller area chart is also shown in figure 1. It contains the exact same information as the bar chart, but it puts into evidence some detail that would otherwise be less marked: for instance, that the peer-assessment is narrow banded in comparison to the tutor's assessment.

Note that the same area plot shows some agreement if we look into other aspects. For instance, the statistical mode matches at 8 for both the lecturer and the peers. Other metrics that seem to agree are the average and the "total marks given" (figure 2) (even though both are slightly larger in the peer-assessment case). The "total marks given", being a summation, can be seen as a measure of the willingness to give better marks. The big difference is on the standard deviation, which is practically twice as much for the tutor. This is because the tutor marks are broad banded in comparison to the peer assessment marks.

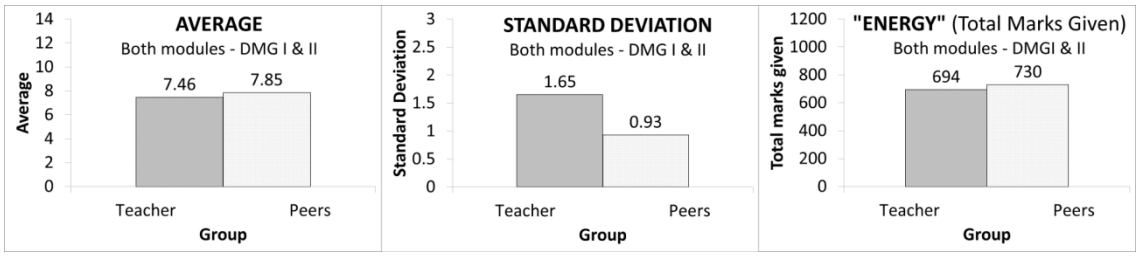

Figure 2. Generic metrics of the results: average, standard deviation and "energy".

It might be expected that, under different circumstances – for example, higher stakes assessments with *category-based criteria* defined - the average and mode would change and shift. In that case, the teacher's assessment would probably be smaller (shifted to the left), as observed by other authors, such as Kirby and Downs (2007) when discussing a case-study on self-assessment.

Qualitative analysis between the lecturer marks and the peer-assessment marks

First, it must be noted that we deliberately refrained from using a quantitative rigorous statistical analysis of the results, as found in some other studies, because we found it might conceal some of the particular causes for the peer-assessment outcomes that we are looking for. Van Zundert et al. (2010) pointed out that most of the published works in the literature about peer-assessment, although providing useful insights regarding best practices, are inconclusive with respect to cause-effect relations involved in the peer-assessment process.

Results on figure 1 suggest that the data can be divided into 2 main groups. We decided to categorise them into "divergent" and "convergent", as we think these words are the most adequate to define how the perception differs from the tutor to the student. The divergent group, in this study, is proposed to be composed by the grades ranging from 1 to 6 (and possibly 10), with zero correlation between the lecturer and the peers,

as opposed to the convergent group ranging from 7 to 10 (or 7 to 9) where data seems to present a certain degree of correlation. Divergence occurs in the 1 to 6 marks range as it sounds evident that the 6 students receiving a pass from their peers (5 and 6 marks) are the same 6 receiving a fail from the lecturer (1 to 4 marks).

To explain what causes divergence between the tutor and the students, selected parts of the plot represented on figure 1 are discussed below:

- Range 1 – 4: This is the range of marks corresponding to a fail. According to the lecturer, 6 students fell into this category. However, the students appear to have avoided failing their peers and did not place any into this category, even though some presentations were quite poor. These results are divergent.
- Range 5 – 6: All the 6 students who received a fail from the lecturer (mentioned in the category above) were put in this category by their peers, i.e. not a fail. Nevertheless, the students demonstrate the ability to rank the presentations: the two students that got 1 and 3 from the tutor were given a 5 (pass) by their peers and the four students that got 4 from the tutor were given a 6 by their peers (slightly more than a pass). These results are divergent.
- Range 7 – 8: 72% of the marks given by the students in peer-assessment fall into this range while 50% of the marks given by the lecturer fall into this category. These results can be said convergent.
- Range 9 – 10: There seems to be a tendency to invert the situation for the highest marks: the students appeared less likely to give a 10 (6 given by the lecturer against only 2 in peer-assessment). Considering the “9” marks, an equal number of “9”’s were awarded by students and tutors. These results can be said convergent, but with a tendency to diverge.

Concerning the first two ranges above, in which the results are divergent between the lecturer and the students, peer-assessment was found to be more generous than the tutor's. However, it is should be noted that the students are able to grade the presentations according to their merit - exactly as perceived by the tutors. All concur that there are 6 presentations worse than the others, in which 2 are clearly worse than the other 4, and this was not a subjective judgement: there was unanimous agreement on this.

On the other hand, the last range (9 to 10 marks) suggests that students seem to avoid marking their peers much higher than what they feel they are likely to get themselves.

Another relevant aspect that is generally pointed out as a source of bias is hostility (Falchikov 2003). However, when dealing with a large number of students and admitting hostility exists only between a few students, this will be averaged out eventually.

What is left is a bulk of marks allocated by the students in the tight and narrow 7-8 range, even though we must recognise the tutors also felt that the majority of presentations would fit in this category.

Normalisation of the peer-assessment marks

One important conclusion from the previous analysis is that the peer-assessment marks are narrow-banded when compared to the tutor marks. Peer-assessment marks range between 5 and 10 (6 marking intervals considered) whereas tutor marks range between 1 and 10 (10 marking intervals considered). In an attempt to stretch this band (i.e., to increase its range) so as to match the tutors' marking range, the following formula is suggested:

$$s_{norm,i} = r \cdot s_i - c \tag{1}$$

In this equation, $s_{norm,i}$ is the normalised peer-assessment mark for student *i*, s_i is the original peer-assessment mark for student *i*, *c* is a constant used to limit the maximum marks given and *r* is the range correction factor given by.

$$r = \frac{\Delta T + 1}{\Delta S + 1} \tag{2}$$

with ΔT the tutors' marking range ($\Delta T = T_{\max} - T_{\min}$) and ΔS the students' marking range ($\Delta S = S_{\max} - S_{\min}$). In this case, $S_{\max} = T_{\max} = 10$, $T_{\min} = 1$ and $S_{\min} = 5$.

Determination of c can be done by imposing one chosen criterion, for instance the maximum mark allowed or the target average. In this case, the criterion adopted was that the average would be the same for both the students and the tutor, i.e.:

$$c = \frac{1}{N} \sum_{i=1}^{N} \left(r \cdot s_i - t_i \right) \tag{3}$$

The values of s_i and t_i are the marks received by student *i* from their peers and the tutor, respectively. These values are the exact averaged values and not rounded values as shown in the histogram earlier. In the present study, the values so obtained were $r = 1.667$ and $c = 5.564$.

The correlation between two sets of marks, say the peer-assessment marks and the tutor's marks, can then be determined from the *Pearson product-moment correlation coefficient*:

$$\rho_{s,t} = \frac{E\left[\left(s - \bar{s}\right)\left(t - \bar{t}\right)\right]}{\sigma_s \sigma_t} \tag{4}$$

where $E[.]$ denotes the ensemble average value of [.] and σ_s and σ_t represent the standard deviation of the peer-assessment marks and tutor's marks respectively.

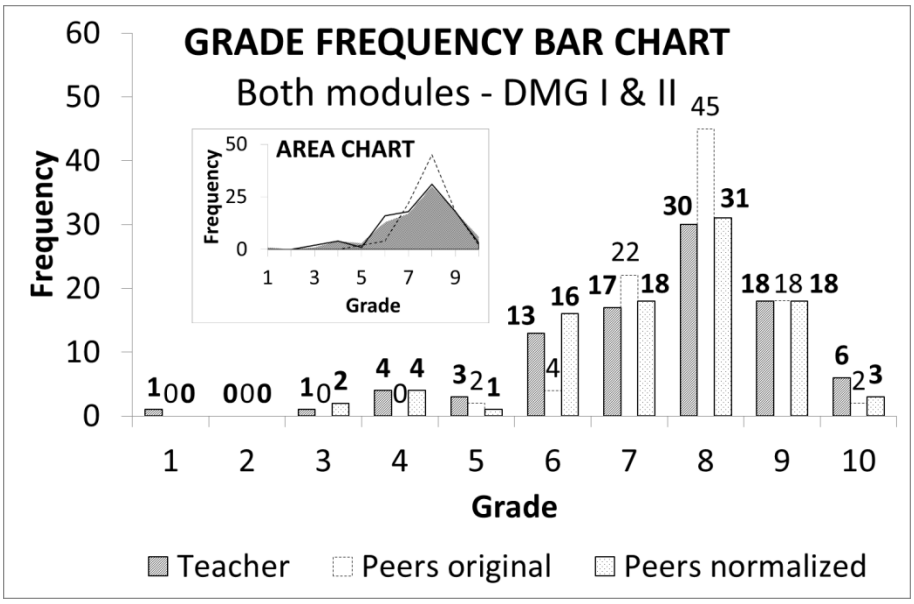

Figure 3. Histogram showing the amount of times a grade was given, including normalisation of the peer-assessment marks.

Results of the normalisation procedure are shown in Figure 3 bar chart, along with the tutors' marks and the original peer-assessment marks. It is very interesting to notice how this adjustment process changed data. Now there is a very strong correlation between the two sets of data (tutor and peers-normalised), highlighted by the area charts that are almost perfect matches. In terms of the Pearson correlation factor between the peer and tutor marks (equation 4), it improved from 94% to 99% after normalisation. Analysis of these results will be done following the same reasoning as before, i.e. by highlighting selected ranges of the plot represented on figure 3:

- Range 1 – 4: The same 6 students that were identified by both the students and the tutors as having the worst presentations are now in the 'fail' range (peer-

assessment originally placed them at the 5-6 range). However, even though the normalised range was stretched so as to include 1, the worst mark obtained was as high as a 3. It is interesting to note that, now, only one student is divergent between the tutors' perception and the students' perception.

- Range 5 – 8: All results are now very similar when comparing the tutors and the normalised peer-assessment marks, especially with respect to the 6th and the 8th marking ranges, in which the relative difference between the number of students included by the tutors and the peers drops from 69% to 23% in case of the 6th range and drops from 50% to only 3% in case of the 8th range.
- Range 9 – 10: There was some improvement in this range as well. This normalisation allowed that one student being graded 9 by their peers now receives a 10 and one student being graded 8 by their peers now receives a 9.

Finally, a comment on the apparent coincidence between the lecturer's marks and the normalised peer-assessment marks. At a first glance, it may seem that equation (1) and the subsequent reasoning serves only to adjust the numbers so as to fit the tutor's marks. Nevertheless, this is only valid happen because the qualitative judgement of the students is coherent with the tutor's judgement. What is happening is that the different presentations were naturally sorted from the worst to the best because the criteria (being *holistic*) tended to be based on comparison.

Conclusion

In this study, an attempt has been made to better understand how well peer- and tutor assessment correlate, even if the awarded marks differ significantly initially. Engineering undergraduate students were invited to get involved in peer-assessment by marking one another's presentations about their Technical Drawing and Geometrical

Modelling projects. A set of marking data with statistical significance (93 students' sample) was analysed to assess how differently the students mark their peers when compared to the tutors.

The peer-assessment was *blind* – the marks were kept anonymous from the students - to reduce any fears of discrimination. However, it was not possible to implement it in a double-blind peer-review fashion, as the students knew who they were marking, so it is possible that fear for discrimination and hostility were not totally reduced, although hostility is expected to be very localised and, thus, averaged out.

As might be expected, students tend not to fail their peers – neither do they tend to award the highest marks either. This means that the band where the students give their marks tends to be narrower than the one used by the lecturer. Yet, it was found that they are able to make a correct relative judgement between their peers' performances (and thus themselves).

To mitigate this, a normalisation procedure to adjust peer-assessment marks has been presented, in an attempt to correlate peer- and tutor assessment in quantitative judgements. This adjustment, that proved to significantly improve the correlation between the tutor's and the students' assessments, is still very dependent on the teacher's perception (e.g., marking range and average). Thus, further study is still needed so as to conclude about its practical usefulness, especially when tutor's marks are not available to establish reliable normalisation parameters.

References

Boud, D., and Falchikov, N. (2006) Aligning assessment with long term learning. *Assessment and Evaluation in Higher Education*, 31(4), 399–413.

Brown, S., Sambell, K., and McDowell, L. (1998) What do students think about peer assessment? In Peer Assessment in Practice, (ed) Brown, S., 107–112. Birmingham, UK: Staff and Educational Development Association (SEDA), paper 102.

Chuck, J.-A., and Young, L. (2004) A Cohort-Driven Assessment Task for Scientific Report Writing. *Journal of Science Education and Technology*, 13(3), 367-376.

Cotner, S. H., Fall, B. A., and Wick, S. M. (2008) Rapid Feedback Assessment Methods: Can We Improve Engagement and Preparation for Exams in Large-enrollment Courses? *Journal of Science Education and Technology*, 17, 437-443.

Dochy, F. J. R. C., and McDowell, L. (1997) Assessment as a tool for learning. *Studies in Educational Evaluation*, 23, 279-298.

Earl, L. M. (2003) *Assessment as Learning*. California: Corwin Press Inc.

Falchikov, N., and Boud, D. (1989) Student self-assessment in higher education: a meta-analysis. *Review of Educational Research*, 59(4), 431–470.

Falchikov, V. (2003) Involving students in assessment. *Psychology Learning and Teaching*, 3(2), 102–108.

Gipps, C. V. (1994) *Beyond Testing: Towards a Theory of Educational Assessment*. London: Falmer Press.

Hatzipanagos, S., and Rochon, R. (2010) Introduction to the special issue: approaches to assessment that enhance learning. *Assessment and Evaluation in Higher Education*, 35(3), 491–492.

Hunter, D. (1999) Developing peer-learning programmes in music. *British Journal of Music Education*, 16, 51–63.

Kirby, N. F., and Downs, C. T. (2007) Self-assessment and the disadvantaged student: potential for encouraging self-regulated learning? *Assessment and Evaluation in Higher Education*, 32(4), 475–494.

Norton, L. (2004) Using assessment criteria as learning criteria: a case study in psychology. *Assessment and Evaluation in Higher Education*, 29(6), 687–702.

Orsmond, P. (2011) *Self- and Peer-Assessment: Guidance in Practice in Biosciences*. Leeds: UK Centre for Bioscience, The Higher Education Academy.

Russel, M., Haritos, G., Combes, A. Individualising students' scores using bling and holistic peer assessment. *Engineering Education*, 1, 50-60.

Silva, J. M. M., and Maia, N. M. M. (1997) *Theoretical and Experimental Modal Analysis*. Somerset, UK: Research Studies Press Ltd.

Tsivitanidou, O. E., Zacharia, Z. C., and Hovardas, T. (2011) Investigating secondary school students' unmediated peer assessment skils. *Learning and Instruction*, 21, 506-519.

Van Gennip, N. A. E., Segers, M. S. R., and Tillema, H. H. (2010) Peer assessment as a collaborative learning activity: The role of interpersonal variables and conceptions. *Learning and Instruction*, 20, 280-290.

Van Zundert, M., Sluijsmans, D., and Van Merriënboer, J. (2010) Efective peer assessment processes: research findings and future directions. *Learning and Instruction*, 20, 270-279.