Multi-disciplinary Peer-mark Moderation of Group Work

Peter Willmot (Corresponding author)

Wolfson School of Mechanical & Manufacturing Engineering, Loughborough University

Loughborough, Leicestershire, LE11 3TU, UK

Tel: 44-1509-227-555   E-mail: P.Willmot@lboro.ac.uk

Keith Pond

School of Business and Economics, Loughborough University, Loughborough

Leicestershire, LE11 3TU, UK

Tel: 44-1509-223-290   Email: K.Pond@lboro.ac.uk

Abstract

Self and peer assessment offers benefits for enhancing student learning. Peer moderation provides a convenient solution for awarding individual marks in group assignments. This paper provides a significant review of peer-mark moderation, and describes an award winning, web-based tool that was developed in the UK and is now spreading across the world as an open-source web application. It is available for use in any discipline. Qualitative research, at the home institution over several years, reinforces the evaluation of quantitative data extracted from the system and from an extensive user survey to confirm, update and strengthen the previous literature. The research also describes new insights into the thoughts of students, who appear to recognise the transparency that automated moderation offers. The statistics suggest few incidences of team-collusion when entering data, but indicate that peer-marking behaviour is influenced by group size, selection method and year of study. Students comment positively on the recognition of their levels of achievement within a team.

Keywords: Peer assessment, Group work, Project-based learning

1. Introduction

Graduates must be well grounded in appropriate subject knowledge and must also be able to operate efficiently in the world of problem-solving, decision making, and cooperative enquiry (see, for example, Rees et al., 2007). It is well researched that collaborative forms of learning, such as group tasks, can help foster lifelong learning and key transferrable skills (Falchicov, 1988; Brown & Pendlebury, 1992 and Boud et al., 1999). Consequently, learning in groups is commonly seen as an activity that stimulates students and, taking Loughborough University as an example, group work takes place in over 500 modules representing every department of the institution (Blease, 2006).

The difficulties of precise assessment of individuals working in groups are substantial. Academics, who feel comfortable setting examinations and individual assignments, may be deterred from group assessments because this student-centred approach means that there is only limited first-hand knowledge of the real contribution each group member makes (Brown & Knight, 1994). Many different methods of reflecting individual effort in group work have been tried, and one of the most popular is 'peer assessment.'

1.1 Context

The terms '(self and) peer assessment' are used to describe the processes where students assess their (own and) peer group performance in relation to a group task. Where this assessment is used to modify a group mark, already allocated by a project supervisor, it is more accurately called 'peer-moderation'. Peer review is highly regarded and universally

accepted in academic research and peer assessment, though not the same, shares the common desirable element that participants actively engage in the process. Student engagement is a clear issue in UK Higher Education today (THE, 2009). Studies have consistently shown that student-centred group work can be more engaging than traditional lectures (Mann & Robinson, 2009).

Peer mark moderation represents a balance between the responsibilities of faculty staff and of students in the educational process. Students alone cannot deliver successful outcomes; group work must be of good design, with sensible assessment criteria and simple mechanisms, to maximise engagement. Students have been quick to detect any moves to shift the burden of responsibility onto them; for example the student "revolt" against the London School of Economics (LSE) after its Chair appeared to suggest that undergraduates were not "cost effective" (LSE, 2009). At Bristol, too, "independent learning" was seen as "no lectures at all" (de Bruxelles, 2006). To date, none of the problems associated with the perceived shift of responsibility from staff to students have arisen within the mark moderation context that is the focus of this paper.

2. Literature review

There are many claimed benefits of self-and-peer assessment for both teachers and learners: the main potential benefit for teachers is that, to enable legitimate group work can save a huge amount of marking and reduce the ever growing workload: for students, peer moderation has the potential to mark group work more fairly. Involving students in marking is not without its problems, and academics should consider these before proceeding. The most important areas to address are, setting the criteria, forming the groups, making provision for handling and reporting the assessment data and ensuring that the whole process is transparent. Clearly, a quality automated peer assessment system must provide assistance and advice to users at both the setup and reporting stages.

Falchikov (1995) identified two distinct types of peer assessment, of product and of performance (sometimes referred to as 'process'). Peer assessment of product is where students assess other students' delivered work: either a finished product, in case of summative assessment, or a work in progress in the case of formative assessment. Peer assessment of performance is where students assess contributions and attitudes to work. Peer assessment is almost universally focussed on product while peer mark-moderation of groups could conceivably be of product and/or performance as defined by the criteria. Clearly, tutors may choose to assess once or more than once: at the end of the work or at various stages along the road. Various examples show that peer assessment can be used formatively or summatively, with the latter being the most reported in the literature. Robinson (2006) demonstrates an example of summative peer assessment whilst Wheater et al (2005) compare two case studies: one summative and one formative, to show successes with both types.

2.1 Background

In sports, commerce or project management it is often the case that the whole team is affected equally by collective team success or failure. Extrapolating this argument suggests that individuals should be prepared to entrust their future to the shared outcome, where that outcome is the quality of the product and a group mark is awarded. In addition, if the primary learning outcome in a course is to develop team-working skills, this seems a reasonable strategy to adopt. This argument, however, is not one that is easily accepted, either by the students or by teaching-quality assessors (Blease, 2006). The fairness of allocating equal marks to all team members was questioned by Willmot and Crawford (2007) who suggested that this was not the correct approach and stated that the common belief is "a lazy student might benefit from the efforts of team-mates, or particularly diligent students may have their efforts diluted by weaker team members". The terms 'free-riders' and 'stars' are frequently used to describe these students (Roberts et al., 2007). It is intuitive that any given cohort of Higher Education students will display a wide range of contribution levels and team-working skills, and peer moderation can go some way to reflecting this in marks. Unfortunately, tutors or project supervisors cannot be solely relied upon to identify and penalise "free-riders" or to reward "stars" (Brown & Knight, 1994) as it is near impossible for a tutor to assess students' individual efforts in a group task when the majority of work takes place during non-contact periods.

Race (2001) identifies that, "when it comes to measuring an individual's relative contribution to group work, the only people who really know about the relative contributions, are the students themselves". He argued that tutor assessment is not sufficiently valid and that students are better placed to assess each-other's work. By involving students in the assessment, teachers gain an additional insight into group dynamics and can measure things that are not otherwise apparent. The validity of peer assessment has usually been evaluated by surveying participants, and various studies have found that students perceive such assessments to be fair (Pond et al., 2007, Crocket, 2003).

The concept of "fairness", however, is abstract and issues of unfairness can emanate from both tutor-based and student-based marking. As Nordberg (2008) points out, there are formative benefits and learning is enhanced when

group projects are used, even if there is a cost in "fairness" (measured against individual assessments) – but what is the alternative? Fairness of marking in peer moderation of individual contributions may well suffer from self-mark over-estimation (Davies, 2000 and Willmot & Crawford, 2004); unwillingness to "mark down" (Conway et al., 1993) and a "training" effect, whereby repeated use of a peer assessment mechanism reduces variance in marks (Resta et al., 2002).

2.2 Potential benefits

Transparency of criteria and peer-to-peer anonymity are often cited sureties for fairness, (Moreira et al., 2003; Lin et al., 2002 and Pond et al., 2007) but, in addition to providing for fairer assessment, the direct involvement of students in the marking process is recognised as contributing to enhanced learning (Nordberg, 2008).

Peer assessment systems have claimed numerous benefits for enhanced student learning and skills development. Russell et al (2006) explore the potential benefits of group work and identify that peer assessment can improve transferable skills including; "decision making, negotiation, communication, empathy and delegation"; Willey and Gardner (2009) show that self and peer assessment is successful in assisting students to achieve specific learning outcomes, and Falchikov (1995) described improved reflection and higher level thinking. In a wider sense, Boud et al (1999) declared that "assessment is the single most powerful influence on learning in formal courses" and Somervell (1993) recommended that self, peer and collaborative assessment should be part of a process of change towards a more student-centred approach to education. Such a strategic leap highlights the significance of designing assessments that stimulate learning whilst achieving the course aims and objectives. The constructivist approach requires a change in emphasis, from the norm-referenced to the criterion-referenced assessment, from the purely summative to the formative and summative, from external to internal and from the assessment of product only to the assessment of process as well.

2.3 Potential pitfalls

Peer Assessment is not universally embraced: most students and academics regard assessment as an obligation on tutors and that it is their duty to assess students (Venables, 2003; Lin, 2002): peer assessment delegates a large part of this responsibility, although peer moderation retains greater control. There can be unwelcome opportunities for vindictive marking and collusion (Brown & Knight, 1994; Conway et al., 1995). One commonly cited pitfall is failure to prepare students for peer assessment and the need to explain the process properly. Discussions of the criteria beforehand are helpful (Juwah, 2003 and Moreira et al., 2003) but not always easy to initiate and students need to understand how to apply the assessment criteria (Cheng et.al., 1997 and Lin et.al., 2002). Of course, this assumes that the methods employed actually have explicit criteria and indeed, this is not always the case: it is quite common for team members to be asked simply to rate each other at the end of a project through some simple metric, however, this mechanism clearly offers little pedagogic validity. Reliable and valid assessment should measure against specific targets that are aligned to intended learning outcomes and course content. Research into reliability more commonly addresses peer assessment of product rather than of performance, but validity can be tested in both types. Langan & Wheater (2003) report a strong correlation between tutor marks and student marks while some earlier researchers (Kegel-Flom, 1975 cited by Falchikov, 1995) argue that they found insufficient reliability. Many stress the importance of creating an open dialogue with the students when explaining the scheme. Zhang and Johnston (2008) argue that student/teacher congruence is not necessary, since reliability of student ratings can be construed as consistent with each other. Concern has also been raised that the confidence in University assessment practices of 'external' communities, particularly those associated with employment of graduates for professional careers, may be negatively influenced but Langhan and Wheater (2003) dismiss this as of minor relevance if peer-assessment is integrated into a diverse portfolio of assessment, on the understanding that learning outcomes are being achieved.

Pond et al. (2007) highlight the lack of objectivity a student could bring in marking their friends and the influence of personal animosity; this suggests that peer assessment can have a negative effect on students' personal relationships within a group. This problem appears to grow or diminish depending on the actual methods employed (Conway et al., 1993). It is anecdotally reported there is very little variation in marks allocated to individuals where the method requires teams to sit together and agree 'each-others' contribution levels. It seems that students are often afraid to speak up. Indeed, some report that this just serves to increase the number of complaints of unfairness, and this characteristic was demonstrated by Willmot and Crawford in a national workshop for engineering lecturers in 2003 (reported at ICEE 2004) and repeated at EE2008 by Loddington et al. (2008) with similarly striking results.

Another obstacle suggested by Falchikov (1995) is that peer assessment might be time-consuming for students and they would object to this imposition. The time taken clearly depends on the system design and is therefore in the hands of the course leader. Orsmond et al. (1996) also believe that, in comparison to traditional methods, peer assessment can be too demanding of students, too time consuming and criteria setting can be problematic. These are problems that were

addressed by the designers of the online tool. Notwithstanding all these issues, most authors note general student acceptance of the methodology.

3. The Web-PA online self and peer-moderated marking system.

Web-PA is designed for groups doing assignments that earn an overall group mark. The system was originally for teams undertaking engineering 'design and make' projects but it has grown into a flexible medium for all types of group work. Each student grades their 'team-mates' and their own performance, and this assessment is used to modify the supervisor's collective group mark. The system is currently used by 65 academics in over half the departments of Loughborough University and has been embedded into the university academic quality systems as the recommended mechanism in group work. An open-source variant has been developed and has been adopted in a growing number of other UK and overseas universities. The software, which is easily customised to suit all subject disciplines, incorporates a number of significant enhancements to help integrate good practice and benefit lifelong learning. The system continues to evolve, taking feedback from an active user-group and its development was described by Loddington et al. (2009). In May 2008, the Web-PA project won an IMS global learning impact bronze award in Austin, Texas. Visitors can access a discussion forum, a demonstrator and numerous support resources at www.webpa.ac.uk

Web-PA was originally developed from a paper-based system to make data-entry and analysis more administratively convenient. It is flexible regarding team size, the number of assessment criteria and the rating scales used by students. During setup, the tutor selects teams directly from their central university database and defines the timeframe within which the students must respond. Students visit the website between the specified dates, and complete a very simple form using clickable radio-buttons. Critically, the tutor who sets the group-learning assignment defines the criteria for each assessment and has the option to add prompts that characterise the intended meaning of each score. For example; a criterion for a group research project might be 'the individual's ability to search and retrieve information: score 1 for 'demonstrated no ability beyond an elementary keyword search; score 5 for 'researched a wide range of sources and found important information.' The system provides helpful advice on how to construct appropriate criteria which a can be designed to assess product, process or both. The assessment may be applied at the end of a project or at any time during it, more than once if required. Data-entry is confidential and only entry points for their own team appear on the student screen.

The software calculates a variation factor for each team member (PA-factor) based on the total scores received for an individual divided by the normalised average scores for the whole team. The tutor or supervisor marks the team submission in the usual way and this mark, or part of it at the supervisor's discretion, is multiplied by the PA-factor for each individual. If all team members score equally, the PA-factor is 1.0 so all members get the team mark but where individuals gains enhanced marks (PA>1.0), other weaker team-members will have their marks correspondingly depressed (PA<1.0). After the deadline, the tutor can retrieve data-reports in a variety of customisable formats and retains the option of intervening if foul play is suspected.

3.1 Research questions

Over a number of years discrete research projects have combined to focus on the evolving needs of both pedagogy and system design. This paper summarises the research and findings from a number of projects. The key research questions were as follows:

- Is online peer mark moderation considered by users to be fair?
- Do teams conspire to achieve particular results?
- Are 'free-riders' and 'stars' in teams adequately dealt with or rewarded?
- Are the results affected by external influences such as how the teams were formed, gender and year of study or subject discipline?
- How can feedback from the system be more effective?

3.2 Research methodology

The research benefited from having unique access to copious good quality data captured by the system that also provided convenient email links to students such that surveys could be sent out to a large number studying a range of degree subjects. In 2005/6, a Higher Education Authority (HEA) "Small Grants make a difference" fund provided research funding for a number of focus groups, which produced some high quality insights into the Peer Assessment process from Business School students. The focus groups aided the design of a wider student survey in 2007. An additional 'Academic Practice' grant in 2006, awarded jointly to the present authors, concentrated on quantitative

analysis of the Web-PA data and prompted further qualitative research into staff and student interactions with the system and advised modification and upgrades to the software platform.

Staff interviews were used to help formulate the survey. In the first semester of 2006/07, Web-PA was limited to a small number of Loughborough staff and so the first interviews were also limited but they reflected the views of 'champions' and 'early adopters'. Repeating the interviews later would have provided a much wider population. The interviews were used primarily to aid focus of the survey questions – much as in the earlier focus groups.

A Student survey was carried out in 2007 using a commercial online survey tool. Providing for a 'prize-draw' inducement, the survey was sent out to 2209 students studying on 36 modules in 14 departments. There was ultimately an overall response rate of 13% with 284 usable responses. The survey had 27 Lickert scale questions and a number of static data questions such as department, year and gender. The Lickert scale questions interrogated the friendliness and ease of use of the system, and if peer assessment had helped in the development of team-working skills. A number of questions focussed on perceptions of 'fairness' with respect to the system's ability to identify strong and weak team members and the students' own feelings about the accuracy of their final mark and whether or not they would welcome more feedback about how it had been derived. A final set of questions asked whether teams had conspired to 'fix' their mark profile and interrogated the effect of 'anonymity': students' willingness to reveal their peer marks to other team members and their desire to know others' marks.

The next part of the research was to analyse the raw Web-PA data captured by the system in 2008. Data were collected from 6 modules across 3 departments, and included group assignments taken by students in all years, both undergraduate and postgraduate: the data-set reflects 730 student interactions. The analysis searched for numerical evidence of fairness/unfairness and honesty/ dishonesty of marking and, again, on collusion within teams.

Finally, current practice-based research activity, supported by the UK Higher Education Academy (Pond, 2010), has focussed on enhancing learning through feedback, and the latest versions make provision for students to add text comments to their scores and to receive their own feedback through a simple anonymised précis of their performance as perceived by their peers. These features are software selected when required by the tutor the results include brief findings from five groups of business studies students who used this in 2008 and 2009, three groups at Loughborough and two in Singapore.

3.3 Key findings of the student survey

Loughborough has a particularly large engineering faculty and it was from here that Web-PA originated so it is perhaps not surprising that there was a numerical gender bias towards males in the survey (62%). Figure 1 demonstrates the breadth of the survey and names all departments that provided at least 5% of the total response. The majority of the groups were found to be formed by tutor-selection while a significant minority had been formed by the students themselves. It can be assumed that in these cases the students knew each other before the start. The residue (14%) used the method of team-formation known as 'seeding': formed by students around a seed-member predetermined by the tutor to have specific attributes.

<figure 1 should be placed here>

Figure 1. Breakdown of respondents by Department

Standard statistical analysis tools have been used to analyse this significant survey. A broad discussion of the key findings follows. Using an ANOVA analysis of mean scores, we identified the most significant and statistically most reliable responses. Note: The specific questions analysed in each section are summarised in the tables and the results highlighted in bold are further discussed.

Table 1. Mean responses to survey questions on perceptions of fairness & anonymity by year-group

<Table 1 should be placed here>

The quantitative data, presented in tables 1 and 2 are based on a 5-point scale where 5.0 represents total agreement, 1.0 represents total disagreement and 3.0 shows neutrality. The average score and standard deviations are recorded for N respondents to each question. Table 1 shows how much postgraduate students differ from undergraduates in considering the own peer reviewed mark too generous (F=5.39, p=0.001). The system appears to be firmly accepted as a means of identifying weaker students (F=3.31, p=0.021), however, final year undergraduates show more reservations of the system's value in this respect and this could be related to their more 'individualistic' approach to team work because of their sharp focus on a final overall grade. Table 1 also suggests that postgraduate students differ from undergraduate

students regarding the disclosure of their marks to their peers (F=3.22, p=0.23). Postgraduates, in fact, would prefer that other group members could know what marks they had given them.

The gender-specific results indicate that women tend to value the anonymity that the present system offers more strongly than men do; this difference is highlighted by a high level of significance (p < .005). They also tend to value the importance of Web-PA for understanding their role within the team. On the other hand, men seem more characterised by a sense of 'camaraderie': they reported to have found it more difficult to give to own friends in their group a low mark, even when this was deserved. The results by gender are summarised in table 2.

Table 2. Gender differences

<Table 2 should be placed here>

Through the survey, we measured attitudes to feedback, i.e. opinions on feeding back the assessment of the peer group to an individual's work. Clearly, this is a sensitive area that requires careful treatment but it is within this concept that there is basis for the frequently heard claim that peer review can develop key skills. While the analysis showed that not much difference exists between the various departments in the sample, there was significant difference of opinion regarding the desirability of offering anonymised feedback on the marks that a student is given by his or her peers against the various criteria. This is an optional routine for this exists in Web-PA. While postgraduate students of all departments appear keen to share their inter-group feedback and undergraduate students of some departments like Business School/Economics and Politics/International Relations appear to appreciate such feedback; students from English & Drama and Engineering would prefer not to disclose or receive any indicator of the peer-marks. This suggests that, in the case of peer assessment, "one size does not fit all" and points the way to a necessary for flexibility in setting peer assessments.

3.4 Quantitative data analysis

The data for this section were extracted directly from the online system. Six modules from academic year 2007/2008 were selected to encompass as wide a variety as possible; the authors considered how 'honest' marking might look from a data point of view.

3.5 Assessing honesty

Honest' marking implies there is a willingness to discriminate between team members and we would expect there to be engagement with the process. So for 'honest' marking there will be a good chance of a student marking him/herself lower than others in the team: **'self-mark < peer-mark'** (column H, table 3 shows the proportion of the cohort to which this applied). The opposite would be where a student seriously overestimates his/her own scores. We would not expect the **groups to give out 100% of all available marks as no group is perfect (**column I, table 3 shows the proportion of groups that awarded more than 95% of all available marks and hence were considered to have failed this test). A **null return, or non participation in the assessment** suggests a lack of engagement **(**column F, table 3 shows the number of students who failed to submit data). Furthermore, one would expect real variation in the performance of any group of individuals against a range of criteria, so a very low or zero standard deviation across the group would indicate possible collusion, laziness or a reluctance to take the review process seriously. This was expressed numerically as a **low percentage of zero standard deviations** (column G, table 3) by comparing the standard deviation for individuals within groups with the standard deviations of the whole cohort. There were hardly any instances of zero standard deviation or null return.

The data collected show that sometimes there is honesty as defined above, and sometimes not. For all the groups, there is a reasonable chance of a 'self-lower than peer' score, but there is variation in the other three measures. Interviews with staff and the survey of student users suggested that method of group selection, year group and group size all have an effect on 'honesty'. To test 'honesty', data were assembled for six modules; those with the largest population (from 3-departments) and we recorded variations with respect to method of team formation method and year of study. For all the groups, we found a reasonable chance of a 'self less than peer' score, but there is variation in the other three measures (above).

Table 3. Summary of analysis

<Table 3 should be placed here>

The self selected group, module 5, had a much higher rate of non-submission and a much higher proportion of 'low standard deviations' than the seeded but otherwise similar year-2 module (6). More than 40% of the self-selected teams had also allocated almost all of the available marks (>95%). In contrast, the 'seeded' groups fared much better with all three tests, scoring under 5%, which combine to suggest less honesty in the self selecting groups.

Taking modules-1 and 4 together, both of which are from the first year and of similar size, the key difference is that, in module-1, the tutor has allocated teams randomly and module-4 is seeded. Both of these modules appear to demonstrate 'honest' marking, with the seeded module having a particularly low percentage of zero standard deviations. Considering module-2 and module-3, (both third year and similar group size) the key difference is that module-2 is random and module-3 is seeded. These modules also exhibit similar marking behaviour but would not fit the criteria as 'honest'. So we can speculate there is no apparent difference between marking behaviour for random and seeded groups, whether that behaviour is honest or not.

When comparing two modules that were similar in every respect except 'year of study' (modules-1 and 2) the tests suggest much less honesty in the year-3 module compared with that from year-1. Both modules were tutor-selected, so behaviour may have been influenced by the fact that third year students are likely to know each other well and behave like a 'self selected' group.

Of particular interest, were two modules from years-1 and 2; both seeded and of similar size: numbers 4 and 6 in the table. These groups both exhibited 'honest' marking behaviour with particularly low zero standard deviations. Students on the year-2 module had experienced the peer review process in their first year and appear to have confidence in and commitment to the process. However, an alternative explanation might be the style of introduction to the process that this lecturer uses as both modules had the same Responsible Examiner. In summary, analyses of the data suggest:

- Self-selecting groups are less discriminating and potentially less 'honest' in their marking.
- Early years students show more 'honesty' than finalists do.
- Finalists show a greater number of zero standard deviations in marks at group level, which suggests more harmonious teamwork.

3.6 Closing the loop by providing feedback

Though pedagogically desirable, not all students require or welcome feedback; even though they were invited to read feedback from the system, many chose not to do so. Anecdotally, willingness to access and note feedback may be related to student expectations. Where marks fall outside expected levels, students tend to seek feedback; where mark levels are as expected, assignments are sometimes not even collected from the lecturer.

As a first step in designing effective feedback for students using Web-PA, and thereby focusing attention on collaborative peer learning, rather than the product of the group work (Willey & Gardner, 2009), selected cohorts of students were given the opportunity to justify and support their numerical peer scores with textual comments, giving an authentic "student voice".

The 'text feedback' option was, at that time, part of a pilot system in use at Loughborough, which has now been incorporated as a software option. There was a noticeable difference in participation rates when business students from the UK and from Singapore were invited to add comments within Web-PA. Students from Singapore were more inclined to leave text comments and this is shown in Table 4. This may indicate a fundamental cultural difference, however, it could be explained by the fact that these UK students were altogether more familiar with the peer assessment process and now accept it as a normal part of their experience. It was also noticeable that students were marginally more willing to post comments about peers than about themselves (shown in the table as the percentage of text opportunities).

Table 4. Uptake of Text Feedback Opportunities
<Table 4 should be placed about here>

Of those students who wrote text, most wrote positive comments. There was a general tendency to write more comprehensively about themselves and to restrict the more negative comments to comments about their peers. Few comments introduced assessment criteria outside those set by the lecturer within the system.

4. Discussion

There is considerable anecdotal evidence that concerns of unfair marking sometimes overshadows the benefits of group work, especially where students are allocated a single team-mark. Peer-moderated marking is an instance of peer assessment that has been successfully applied here, to group projects and assignments, using a particular online tool. The tool can be used in any discipline and the present research did not uncover any quantifiable differences between the results derived from engineering or from business studies. Furthermore, it is known that Web-PA has been successfully used by academics at the home institution and beyond across a wide range of subject areas. The idea of peer mark moderation is not new, but its intensity of use and support for its pedagogic validity appears to be growing.

Web-PA is designed to be extremely flexible in many aspects of its deployment, including group membership selection, frequency of use, formative or summative use, criteria for assessment and the weighting the moderation factor has on

tutor derived group marks. While the software is available free of charge as an open-source product and is customisable to suit any specific instance, it is essential that installation is undertaken by an IT specialist, who is able to provide the appropriate links to the institution's central computer systems. As a non-commercial product, technical backup is not available, however, there is an active, inter-university Special Interest Group (SIG) that can offer advice and share experiences and it is through this that development has continued beyond the initial development project.

Overall, mark moderation is found to be credible, and while 'free-riders' are known to mark themselves up, the overall system compensates for this and generates an acceptable, lower than team-average, grade. Final year undergraduate students, however, exhibit an individualistic approach to study that is heavily focussed on maximising their own mark rather than on any developmental benefits. Institutions planning to use Web-PA for undergraduate students need to manage the level of exposure their students have to it across the programme as a whole, and design its deployment for use with the specific learning outcomes to minimise finalist bias.

While most users seek a simple solution to providing fair and equitable group marks, others have focussed on the potential for peer learning through feedback. Simple numerical feedback, comparing individual scores for each criterion with the group average, is an option built into the online system. This proves to be a useful mechanism to aid skills development whilst retaining anonymity but students do not always effectively reflect on the feedback with the level of maturity required.

Text based peer feedback, is another optional feature that is presently enabled by only a small minority of users. While potentially very useful, this too is not universally welcomed by students and can cause friction unless handled with sensitivity. Maintaining confidentiality is essential. The implications for removing anonymity, even where data protection issues could be resolved, may mean less candour in peer marking and would devalue peer assessment as a tool. Text based comments can also be perused by tutors to substantiate marking of outliers or apparently rogue results. For example, where a free-rider has been marked down by peers, it is not always apparent to tutors why the low marks have been given. Consistent comments by peers about non-attendance or inactivity within a group would serve to justify the marks. A larger study, concentrating specifically on feedback aspects, would be required to identify with which groups the educational benefits of providing a broader explanation of the mark moderation outweighs the increased potential for acrimony.

5. Conclusions

The Web-PA online mark-moderation method provides efficient and effective support for academics who want to move away from the didactic model. It has met with a very enthusiastic and growing following and is currently being used in over twenty institutions worldwide.

Whilst much of the data, presented here, supports previous literature on this subject, important new insights are gained into the thoughts of the student participants. There is substantial evidence, gathered over several years, that Web-PA users perceive it as fair. More specifically, they comment positively on qualities of anonymity, recognition of 'stars' and 'free-riders' and point to a, perhaps surprising, lack of collusion. The implications for practice are that institutions can be confident in using online peer mark moderation in their assessment strategy as a default or recommended mechanism to regulate the use of group work in the curriculum.

Academic users must exercise transparency in its deployment and must manage student expectations from the outset. There are detectable differences in the peer review data according to how the teams are originally constructed but there is insufficient evidence to offer concrete conclusions, except to note that self-selecting groups appear to generate a smaller variance in the marks; this could be because they work better together.

Whilst maximising the opportunities to develop team-working skills and to aid student reflection, tutors must be mindful of the risks to credibility and openness when particular gender balance, group selection method and level of study characteristics combine. While there is strong agreement, throughout, that anonymous marking is preferred, the research suggests women are more strongly opposed to full disclosure of marks than men are, while male students encounter more difficulty in marking their friends. Females are also statistically more inclined to allocate a larger spectrum of marks. In addition, the generally more mature postgraduate students show a stronger appreciation for peer assessment as an educational support tool for developing and refining their own team-working skills, while early years students have little regard for this aspect and finalist undergraduates are particularly protective of their final grade.

Where the text feedback option was implemented, the majority of comments were positive and focused on the assessment criteria within the system. While students appear content to criticise others, they are less prepared to reflect negatively on their own capabilities. The text tool is still in its infancy and has potential, not only to inform feedback to students in the system, but also to give tutors confidence in the assessment.

Online peer mark moderation is a mechanism that saves tutor time and imposes a discipline on students, most importantly, it can be quick and easy to use, criteria driven and confidentiality is assured. It does not offer a complete solution to the less well behaved students but can provide good evidence to support tutor intervention in group disputes. It has certainly proven to be vastly superior in many respects to the various peer mark adjustment procedures where groups are simply asked to distribute a proportion of the marks amongst themselves.

Institutional policies on group work, moderated by peer mark moderation should be informed by this ongoing research.

<Main body word count = 6263>

References

Boud, D., Cohen, R., & Sampson, J. (1999). Peer learning and assessment. *Assessment and Evaluation in Higher Education,* 24 (4), 413-426. http://dx.doi.org/10.1080/0260293990240405

Blease, D. (2006). The organisation and assessment of group work at Loughborough University, internal report.

Brown, G., & Pendlebury, M. (1992). *Assessing Active Learning*, CVCP/USDU., Sheffield.

Brown, S., & Knight, P. (1994). *Assessing Learners in Higher Education*, London: Kogan Page. ISBN 0 7494 1113 9.

Cheng, W., & Warren, M. (1997). Having second thoughts: Student perceptions before and after a peer assessment exercise. *Studies in Higher Education*. 22 (2), 233-239. http://dx.doi.org/10.1080/03075079712331381064

Coates, H. (2005). The value of student engagement for higher education quality assurance. *Quality in Higher Education*, 11 (1), 25-38. http://dx.doi.org/10.1080/13538320500074915

Conway, R., Kember, D., Sivan, A., & Wu, M. (1993). Peer assessment of an individual's contribution to a group project. *Assessment and Evaluation in Higher Education*, 18 (1). http://dx.doi.org/10.1080/0260293930180104

Crockett, G., & Peter, V. (2003). Peer assessment in a second year macroeconomics unit. Higher Education Academy, Economics Network, Extended Case Study, [Online] Available: http://www.economicsnetwork.ac.uk/showcase/crockett_peer.htm

Davies, P. (2000). Computerized peer assessment. *Innovations in Education and Training International.* 37 (4). http://dx.doi.org/10.1080/135580000750052955

de Bruxelles, S. (2006). We want more lectures, say students. *The Times*, November 2009, [Online] Available: Http://www.timesonline.co.uk/tol/life_and_style/education/student/news/article630752.ece.

Falchicov, N. (1988). Self and Peer Assessment of a Group Project Designed to Promote the Skills of Capability. *Programmed Learning and Educational Technology*, 25 (4), 327-339..

Falchikov, N. (1995). Peer feedback marking: Developing Peer Assessment. *Innovations in Education and Training International*, 32, 175-187. http://dx.doi.org/10.1080/1355800950320212

Juwah, C. (2003), Using peer assessment to develop skills and capabilities. *The Journal of the United States Distance Learning Association*. 17(1) [Online] Available: http://www.usdla.org/html/journal/JAN03_Issue/article04.html

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

Lin, S. S. J., Liu, E. Z., & Yuan, S. M. (2002). Student attitudes toward networked peer assessment: Case studies of undergraduate students and senior high school students, *International Journal of Instructional Media*, .29 (2).

Loddington, S., Pond, K., Wilkinson, N., & Willmot, P. (2009). A Case Study of the Development of WebPA: An Online Peer-Moderated Marking Tool, *British Journal of Educational Technology*, 40 (2), 329-341. http://dx.doi.org/10.1111/j.1467-8535.2008.00922.x

Loddington, S. P., Wilkinson, N. J., Bates, M. R. N., Crawford, A. R. & Willmot, P. (2008). Discover WebPA: An online Self and Peer Assessment System, *Proceedings of EE2008: International Conference in Engineering Education*, HEA Engineering Subject Centre, Loughborough, 7 pp, ISBN 978-1-904804-65-9, [Online] Available: http://www.ee2008.info/index.php/papers/.

LSE. (2009). Undervalued students, LSE:Notforrpofit.org, London School of Economics, February, [online] Available: http://lsenotforprofit.org/?page_id=7.

Mann, S., & Robinson, A. (2009). Boredom in the lecture theatre: investigation into the contributors, moderators and outcomes of boredom amongst university students, *British Educational Research Journal,* 35 (2), 243-258. http://dx.doi.org/10.1080/01411920802042911

McFadden, M., & Munns, G. (2002). Student Engagement and the Social Relations of Pedagogy, *British Journal of Sociology of Education,* 23 (3), 357-366, ISSN: 1469-3346.

Moreira, D. D. A., & Da Silva, E. Q. (2003). A method to increase student interaction using student groups and peer review over the internet, *Education and Information Technologies*, 8 (1)..

Nordberg, D. (2008). Group projects: More learning? Less fair? A conundrum in assessing postgraduate business education, *Assessment and Evaluation in Higher Education*, .33 (5), 481-492, http://dx.doi.org/10.1080/02602930701698835

Orsmond, P., Merry, S. & Reiling K. (1996). The importance of marking criteria in the use of peer assessment, *Assessment and Evaluation in Higher Education,* 21. . http://dx.doi.org/10.1080/0260293960210304

Perry, R., & Maiden, B. (2008). Assessed Group work: dealing with "Free-Riders", BMAF annual conference: The Learning and Teaching Agenda in the UK: National Perspectives but Common Concerns? Business, Management, Accountancy and Finance subject centre, Edinburgh, April 2008.

Pond, K., Coates, D. S. & Palermo, O. (2007). Student Experiences of Peer Review Marking of Team Projects, *International Journal of Management Education*, 6 (2), 30-43, ISSN 1472-8117.

Pond, K. Closing the feedback loop in WebPA - Peer Review of individual students in group projects, an 'Evidence Net' briefing paper, Higher Education Academy, July 2010.

Race, P. (2001). Self Peer and Group Assessment,. LTSN (Learning and Teaching Support Network) Generic assessment series, Briefing Paper.

Rees, C., Forbes, P. and Kubler, B. (2007). Student Employability Profiles: a guide for higher education practitioners, The Higher Education Academy,. [online] Available:. http://www2.hull.ac.uk/administration/pdf/Forbes_et%20al_StudentEmployabilityProfiles.pdf

Resta, P., Awalt C and Menchaca M, (2002). Self and Peer Assessment in an on-line collaborative learning enviroinment., In: G.Richards (ed). *Proceedings of World Conference on E-Learning in Corporate, Government, Healthcare and Higher Education,* 2002, Chesapeake VA, pp 682-689, AACE.

Roberts, M., Parker, J., Joyce, J., & Hassall, T. (2007). Improving group work in Accounting: a student perspective, *Teaching Learning & Assessment conference,* Sheffield Hallam University, July.

Robinson, C. L. (2006). Self and peer assessment in group work. The European Society for Engineering Education: SEFI International Conference, MWG Seminar, Kongsberg. Norway,

Rogers, K. M. A. (2009). A preliminary investigation and analysis of student learning style preferences in further and higher education, *Journal of Further and Higher Education*, 33 (1), 13-21.

Russell, M., Haritos, G. & Combes, A. (2006). Individualising students' scores using blind and holistic peer assessment. *Engineering Education: the Journal of the Higher Education Academy Engineering Subject Centre,* 1 (1). 50-59. [Online] Available: http://www.engsc.ac.uk/journal/index.php/ee

Sommervell, H. (1993). Issues in assessment, enterprise and higher education: the case for self, peer and collaborative assessment. *Assessment and Evaluation in Higher Education.* 18 (3), 221-233. http://dx.doi.org/10.1080/0260293930180306

THE, (2009). Rip it up and start again, *Times Higher Education,* 11 June, [Online] Available:. http://www.timeshighereducation.co.uk/story.asp?sectioncode=26&storycode=406915.

Venables, A., & Summit, R. (2003). Enhancing scientific essay writing using peer assessment, *Innovations in Education and Training International,* 40 (3).

Wheater, P., Langan, M., & Dunleavy, P. J. (2005). Students assessing student: case studies on peer assessment. 'Planet'; the *Journal of the Higher Education Academy Subject Centre for Geography and Earth Sciences.* 15, 13-15. [Online] Available: http://www.gees.ac.uk/planet/p15/p15.pdf

Willey, K., & Gardner, A. Changing Student's Perceptions of Self and Peer Assessment. *Proceedings of Research in Engineering Education Symposium (REES2009),* Paper 75, Palm Cove Australia, July 2009. ISBN.1.921047.61.5.

Willey, K., & Gardner, A, (2009). Developing team skills with self and peer assessment: Are benefits inversely related to team function?, *Campus Wide Information Systems*, 26,(5), Emerald, 365-378. [online] Available: http://www.emeraldinsight.com/journals.htm?issn=1065-0741&volume=26&issue=5&articleid=1823925&show=html

Willmot, P., & Crawford, A. R. (2004). Online peer assessed marking of team projects, *Proceedings of the International Conference on Engineering Education, ICEE 2004*, Paper WA7/4, Gainesville, Florida, USA, October, pp 1-7

Willmot, P., & Crawford, A. R., (2007). Peer Review of team marks using a web-based tool: an evaluation, *Engineering Education: the Journal of the Higher Education Academy Engineering Subject Centre*, 2 (1) , 59-66, ISSN 1750-0044, [online] Available: http://www.engsc.ac.uk/journal/index.php/ee

Zhang, B., Johnston, L., & Kilic, G. (2008). Assessing the reliability of self and peer rating in student group work, *Assessment and Evaluation in Higher Education,* 33 (3), 329-340.. http://dx.doi.org/10.1080/02602930701293181

Table 1. Mean responses to survey questions on perceptions of fairness & anonymity by year-group

		FAIRNESS TESTS				ANONYMITY TESTS		
		Ability to identify 'free riders'	Ability to identify 'stars'	Personal mark perceived too Low	Personal mark perceived too High	Want to know others' marks	Happy to reveal marks	Believe anonymous is more accurate
Year-1 (u'grad)	N	159	159	153	153	156	156	156
	Mean	4.77	4.82	3.16	2.3	3.72	3.41	4.51
	Std. Dev.	1.41	1.18	1.46	1.08	1.88	1.81	1.59
Year-2 (u'grad)	N	56	56	55	55	55	55	54
	Mean	4.45	4.66	2.82	2.49	3.38	2.95	4.98
	Std. Dev.	1.57	1.47	1.28	1.07	1.66	1.71	1.44
Year-3 (u'grad)	N	23	23	23	23	23	23	23
	Mean	**3.78**	4.23	3.17	2.17	3.96	4.09	3.91
	Std. Dev.	1.59	1.42	1.72	1.27	1.87	1.67	1.86
Postgraduate	N	10	10	10	10	10	10	10
	Mean	4.70	4.70	3.20	**3.70**	4.40	**4.30**	4.10
	Std. Dev.	1.25	1.42	1.32	1.5	1.65	1.64	2.33
Total	N	248	248	241	241	244	244	243
	Mean	4.6	4.71	3.09	2.39	3.70	3.41	4.54
	Std. Dev.	1.48	1.29	1.44	1.14	1.83	1.8	1.64

Table 2. Gender differences

GENDER TEST		Wanted to know others' marks	Happy to reveal own marks	Believe anonymous is more accurate	Difficult to award marks to friends
Male	N	153	153	152	153
	Mean	3.77	3.55	**4.19**	**3.54**
	Std. Deviation	1.91	1.86	1.79	1.66
Female	N	91	91	91	91
	Mean	3.57	3.18	**5.13**	3.03
	Std. Deviation	1.68	1.66	1.15	1.47

Table 3. Summary of analysis

A	B	C	D	E	F	G	H	I
Module	Method of group selection	Year	No. of students	Average Team size	Non submission %	Zero standard deviations %	'Self' lower than 'Peer' mark %	Teams with >95% available marks
1	Tutor Random	1	286	6.00	7.0%	6.4%	18.8%	2.1%
2	Tutor alphabetical	3	87	3.90	1.2%	40.0%	15.1%	31.8%
3	Seeding	3	69	4.60	4.4%	40.9%	13.6%	26.7%
4	Seeding	1	109	5.70	4.6%	1.9%	27.9%	0.0%
5	Self selecting	2	63	3.20	12.7%	60.0%	18.2%	42.1%
6	Seeding	2	116	4.30	0.9%	2.6%	31.3%	3.7%

Table 4. Uptake of Text Feedback Opportunities

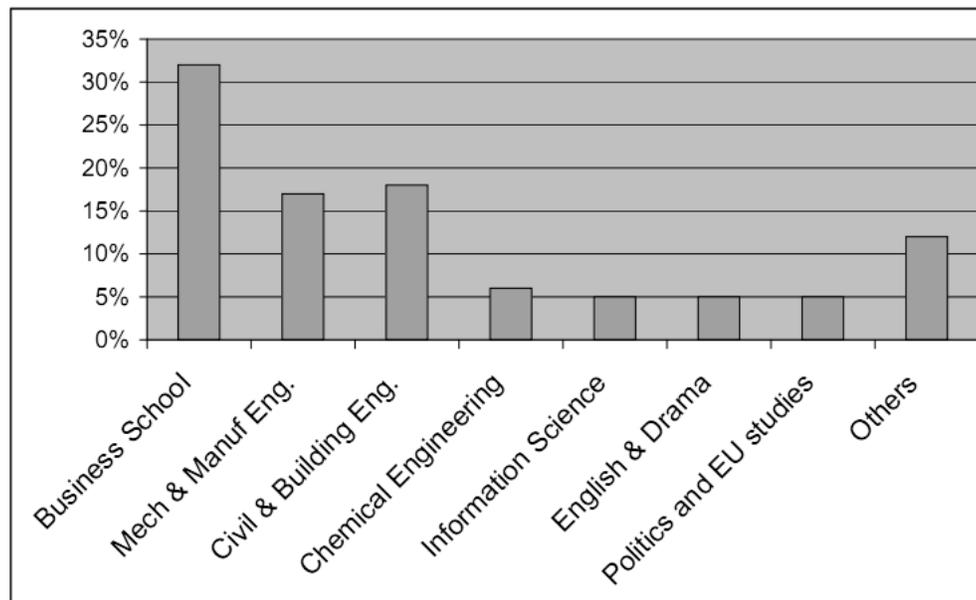| Title and year (all are year 1 modules)	No of students	% text opportunities taken		% text opps.
		Self	Peer	Overall
Singapore				
Personal Effectiveness, 2008	30	71.26%	81.82%	80.77%
Personal Effectiveness, 2009	26	91.67%	92.86%	92.5%
Loughborough				
Banking Securities, 2008	75	56.94%	60.7%	59.85%
Personal Effectiveness, 2008	320	60.83%	61.14%	60.97%
Personal Effectiveness, 2009	268	74.71%	74.17%	74.27%

Figure 1. Breakdown of respondents by Department