Loughborough
University

This item was submitted to Loughborough's Institutional Repository (https://dspace.lboro.ac.uk/) by the author and is made available under the following Creative Commons Licence conditions.

For the full text of this licence, please go to:
http://creativecommons.org/licenses/by-nc-nd/2.5/

Peer review of team marks using a web-based tool: an evaluation

Peter Willmot and Adam Crawford

Abstract

Of all the problems associated with student learning in a team situation, the difficulty of fair assessment of the individual team members is supreme. Academics who feel comfortable setting examinations and single-person assignments are deterred from setting team assessments because they fear that idle students may benefit from the efforts of their team-mates or that weaker team members might dilute the efforts of the more diligent. This paper discusses how accurately academics can recreate the rewards for good or bad performance in industry through undergraduate team projects. The arguments for allocating equal team marks are examined but the authors conclude this is not the correct approach. A web-based system for applying peer moderation to team marks is described and accumulated data from it allows peer marks to be compared with anonymous self-assessments. Validation is completed by comparing the peer assessment outcomes with control data supplied by independent mentors that were attached to each student team. The results generate a high level of confidence in the approach. Peer review results for teams were further used to estimate the degree of harmony amongst team members: a high standard deviation in peer marks might indicate conflict, whereas a low standard deviation could be a sign of a harmonious team that one might expect to out-perform the individual potential of its members. Previous academic track record was used as the benchmark for potential success but was found to be a poor predicator of actual achievement in team project work.

Introduction

Academics who feel comfortable setting examinations and individual coursework assignments are deterred from devising team assessments because they cannot guarantee that they can assess the individuals within the teams accurately. 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 UK Higher Education Academy (HEA) briefing on peer assessment identifies 'the hard part is to assess the group work itself and the relative contribution of individual members of the group to the development of the assessed product' (Race, 2001). Failure to address this very real problem may not only cause resentment amongst the students but is also considered quite unacceptable from the Quality Assurance point of view. Studies by working groups of academics on behalf of a recent joint funded programme on project-based learning based at Nottingham and Loughborough Universities distilled current practice in UK higher education (Crawford et al., 2003). Most acknowledge the problem and that many of the mechanisms in place to address it are less than ideal. Self and peer assessment has been the focus of much of the recent work in this field where marks or weightings, collected from the team members themselves, are used to modify a team mark allocated by the project supervisor through various mechanisms (Lejk et al., 1996). This paper sets out to examine the wisdom or otherwise of applying an automated self/ peer marking method known as Web-PA and to examine some of the pedagogic issues associated with it.

The real world

Some would say that if we are to mimic the 'real world', the emphasis in engineering education should be firmly centred on the team. It might be argued that in industry, just as on the sports field, the outcome is a team effort and all benefit equally from the team's success. When your favourite football team wins the league, for example, even the substitutes receive a winner's medal. Is it fair, therefore, to conclude that individual students must be prepared to entrust their future to the collective outcome in university team assignments that are integral to the modern engineering curriculum. How strong is this argument for giving equal marks to all team players?

Being in a team has become an inescapable feature of modern professional life: we are all 'team players'. In the work context, the perceived wisdom is that teamwork delivers real organisational benefits: improving productivity, reducing absenteeism and enhancing employee satisfaction. The outcome of an effective team is expected to exceed the sum of its component parts. Consider any major engineering or construction project like the design of a new aircraft or the building of the channel tunnel and it is quite clear that no one person has the breadth and depth of knowledge and skill needed for the whole project and this implies teamwork. In the world outside engineering, teams have a clear, highly positive image and are well known to offer the same strength in cooperation. Sporting teams like Manchester United and the New York Yankees are striking examples, as are brass bands, choirs and symphony orchestras in the field of music. A place in such a team confers clear status; where you compete to get into the team and have to perform consistently well if you are to keep your place.

Such examples paint a highly optimistic picture of teamwork, however, and although one may be able to point to some high achieving student teams where all pull together effectively, colleagues in academia will have had very different experiences working with other, less cohesive groups of students. When it comes to work-based project teams, the reality is that project managers don't normally have the luxury of choosing the members – they take who they can get. The same process turns out to be a lottery for industry team members in terms of support and career progression. Some academics argue that project module leaders should take personality profiles like Myers Briggs type indicators or Belbin team roles into account when constructing teams, but most find that this is not really practical and does not in any case reflect what happens in industry. So perhaps we ought not to concern ourselves too much about team selection in degree programmes provided we mix them up now and then. Our experience indicates that an element of self-selection tied in with an element of compulsory intermixing and team rotation is the best route to assembling satisfied teams that represent a 'real world' scenario.

Other differences between a sports team and a project team are obvious: in the role and power of the manager and the importance

of practising. A sports team is managed by someone external to the team, the club coach; whilst a conductor manages an orchestra. These individuals are the arbiters of standards and their decisions are final. Industrial project managers operate in a more complex managerial environment, which inevitably limits their powers and discretion. It is in the area of management control that the student project team moves away from the engineering industry model and tends towards the sports/ music team scenario because academic supervisors are frequently viewed by students as all-powerful. A sports team is, however, given time to practise, and an orchestra to rehearse; project teams are somehow expected to gel and perform from day one. Consequently, is it reasonable to expect our students to hit the ground running at their first experience of teamwork or will they improve with time? Our anecdotal experience shows that students get better at teamwork with practice and that, while teamwork instruction has its place, experience is more effective. This is particularly evident in the speed with which a newly formed team with experienced members goes through its evolutionary stages of 'forming, storming and norming' (Tuckman, 1965) and begins to get down to real work. It is clear that students benefit from tackling various types of teamwork throughout their degree programme.

Clearly, the different types of teams under discussion work in radically dissimilar environments, although it can be argued that it is the project team that is the most complex to analyse. Unlike the football analogy, each project is unique and the conditions for team selection and motivation are often less than ideal. Academics should not feel isolated by this; it is just a matter of degree. As for drawing parallels in academic assessment of individual players, the highly successful sports team will command the greatest accolades and, as already described, even the weakest member of that team will rightly demand a share of the good fortune. Similarly all players in a relegated football club can expect to take a hit. On the other hand, the brightest stars within the cup-winning club can expect further enhanced salaries and advertising contracts while the stars in the relegated club may collect a lifeline transfer. It is similar in industry: a successful project team is judged by that project's success or failure and the strong or weak team members can expect marginal adjustments. This idea sits well with awarding overall team

marks that are later modified by incremental adjustments for individuals, but the questions for academics are how best to make those adjustments and how big they should be.

Web based peer review

An innovative web-based system of self and peer assessment has been in place at Loughborough University across a number of modules and disciplines for several years. It provides a convenient mechanism through which individuals can be assigned a proportion of the supervisor mark based upon data entered by the members of the teams. The system, known as Web-PA was designed to handle large numbers of students in a flexible, tutor driven manner and has been previously described in detail by the authors (Willmot and Crawford, 2004).

Web-PA allows teamwork supervisors to set up a number of criteria against which team members mark themselves (self) and the other team members (peers) at the end of a project or at any other prescribed time within it. Students enter data in confidence from any web terminal and the system calculates a variation factor (Web-PA factor) for each team member based on the total score for an individual divided by the normalised average scores for the whole team. The 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 factor for each individual. Where all team members score equally, the Web-PA factor is 1.0 so all members gain the unmodified team mark. Local experience indicates higher levels of student satisfaction with team assessments since the system was introduced and this is confirmed by experience elsewhere with a similar, albeit paper-based method (Gatfield, 1999).

Independent validation

Clearly it is always going to be difficult to obtain independent validation of Web-PA input data given that the peer review process is employed primarily because knowledge of the precise contributions of each student is lacking. The authors have previously shown that the Web-PA algorithm using confidential data entered against specified criteria compares well with other common types of peer assessments. In the research, the Web-PA output data was also dissected to see how well self marking compared with the marks awarded by the other members of the team (Willmot and Crawford, 2004). In most teams the comparison showed considerable agreement on which members were the main contributors. However, it is perhaps not surprising that an individual's claim for their contribution, truthful or otherwise, is often higher than the group's (peer) estimation. The biggest difference apparently occurs for the weakest group members who probably enter unjustifiably high self marks to avoid failing the module. Nevertheless, the web system was found to consistently and quite correctly punish weak members and reward stronger ones, even though the differences in marks appear less pronounced than the perceived wisdom would suggest they should have been. This does give the assessor a degree of comfort as it means students are seldom actually failed by the application of Web-PA. Furthermore, as discussed earlier, this outcome does appear to accord with the real world scenario that even weak members are justly able to claim some of the rewards for being part of a successful team.

A new approach to validation considers the opinions of level 4 Master of Engineering (MEng) students working with student teams as peer mentors (Figure 1). Second year,

Figure 1. Supervisory structure
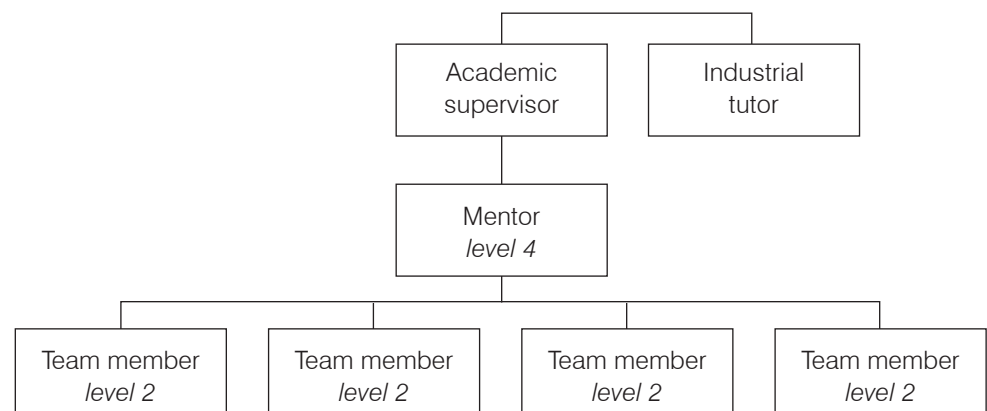
level 2 mechanical engineering undergraduate students at Loughborough routinely tackle a team design project. Students visit an industrial company from a small consortium where they are introduced to a real problem. A finalist 'mentor' works with each team of three or four. Each consortium company sets the problems for four or five teams (16-20 students in all) and an academic supervisor takes charge of the activities of that company group and assessment of the project outcomes, informed by a company tutor. In 2004/5 there were eight company groups operating with more than 30 teams. Projects run from mid October to early May and occupy one afternoon per week. The novel mentoring experience forms part of a 'Project Leadership' module for the finalists which has proved very successful (Willmot, 2003). It appears that the mentors are often more able to effectively communicate ideas than the teaching supervisor who inevitably holds a more authoritarian position. A strong parallel can be seen here between the industrial project team that is mentored by a section leader who is probably not far removed in seniority from the rest of the team but who reports in turn to a senior manager. The senior manager, represented here by the academic supervisor, is usually more detached from the day-to-day work.

Mentors build strong working relationships with their teams through weekly meetings, not all of which are observed by the supervisor, but the age and maturity gap ensures that the relationship remains on a professional level. Mentors focus on the task but are not directly involved in team assessment. They are regularly appraised on their performance as a leader, how they manage the project and

deal with problems that might occur, but their assessment regime is unaffected by the marks of their mentees (the second year students). By the end of the project however, mentors are in an excellent position to offer candid and unbiased opinions as to each student's efforts and abilities.

In the year of the study, mentors were issued with a paper version of the peer review completed by the second year students and were asked to score their team using the same criteria as in Web-PA. Mentors were assured that the information was being gathered for anonymous peer review research. Completing the form was optional and 22 out of 27 mentors responded. The peer assessment process assumes that students can score marks by contributing to any or all of the published assessment criteria and the peer assessment factor is calculated by summing the marks under each heading. The six criteria that the team marked each other on were:

- Ability to find and retrieve technical information.
- Ability to generate ideas and concepts.
- Ability to methodically evaluate concepts.
- Contribution to modelling and development of design solutions.
- Attendance and diligence.
- Communication skills.

Figure 2 is a plot of the mentor response against the weighting factor calculated by Web-PA for the team projects in the academic year 2004/5. The mentor scores were faithfully processed in the same way as in the web tool; that is, mentor assessment equals the score for the individual divided by the average score for all team members. Of the 59 students for

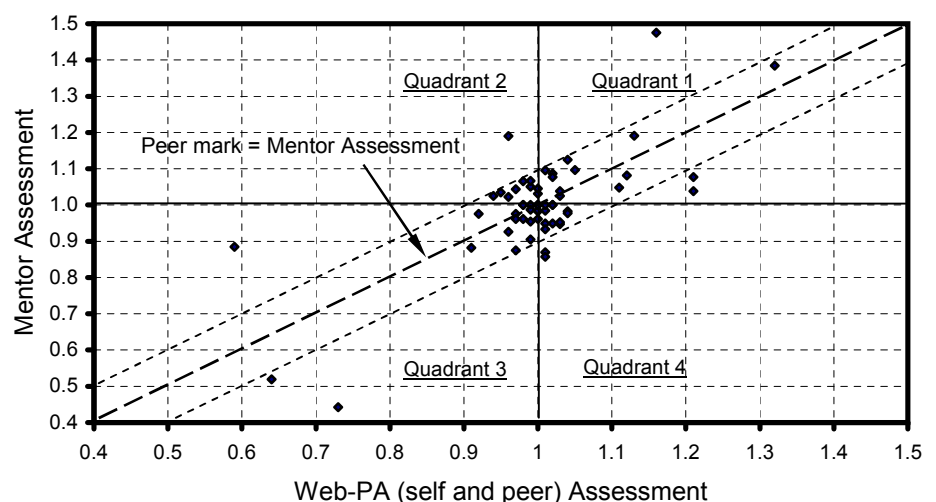**Figure 2.** Comparing Web-PA Output with Mentor Opinions

Figure 3. Average Total Marks awarded to Self vs. Average Marks awarded to other Team Members
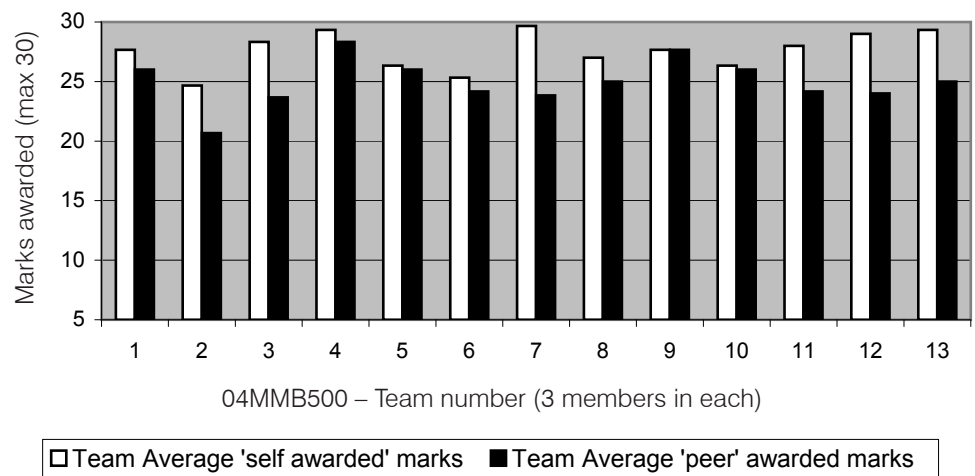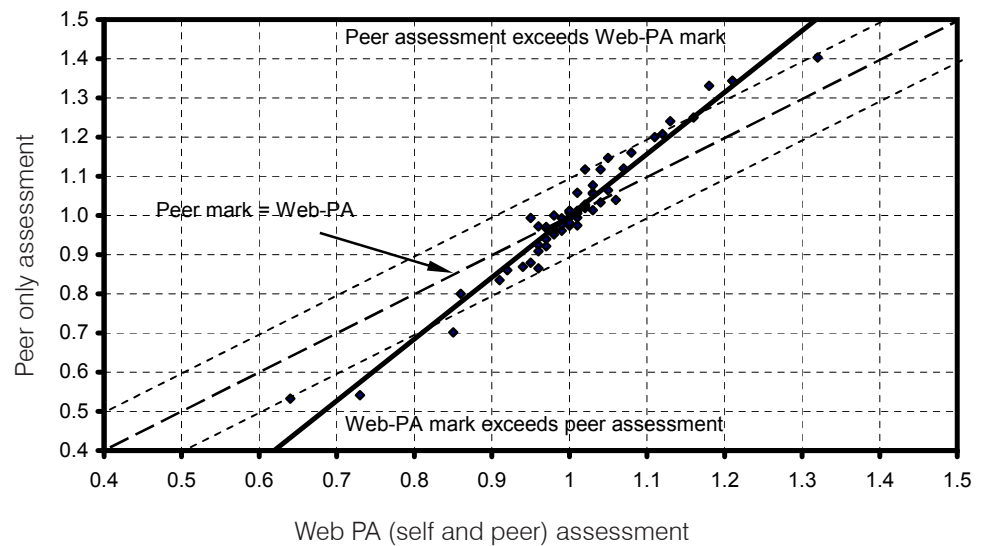

04MMB500 – Team number (3 members in each)

□ Team Average 'self awarded' marks   ■ Team Average 'peer' awarded marks

Figure 4. Peer only Marks vs. Web-PA Scores
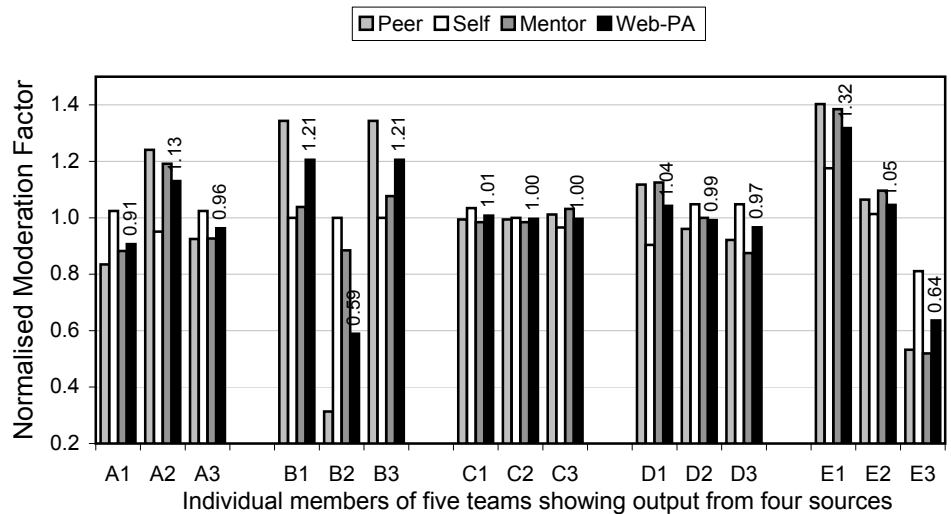

Web PA (self and peer) assessment

which data exists, 85% show broad agreement between the Web-PA result and the opinion of the mentor, defined as within the parallel lines drawn at +or- 0.1 either side of the line of equality. Of the nine students who fell outside these error bars, six are located in the first and third quadrants (both values above unity or both below) which suggest that both response mechanisms agreed that the individual was a strong/weak team member but disagreed about the degree of strength or weakness. Only three students (1.5% of the total) appear in the second and fourth quadrants indicating disagreement.

When one compares the number of marks, on average, that an individual allocates to him or herself with the average number of marks that are awarded to other people (peer marks) the picture is striking. In almost every team examined over two year groups, the average self-mark was higher, confirming that strategic self-marking is widespread. Figure 3 illustrates the trend.

It is interesting to consider the effect of this strategic self-marking on the Web-PA scores by removing the self-assessment marks for the same year group (and including those teams for which no independent mentor assessment was available). Figure 4 shows this output – the opinions of the peer group only – plotted against the normal Web-PA output (which contains both peer and self assessment). A clear trend is discernable: the Web-PA system has the effect of moderating the extremes. Strong students still receive added marks but less so than would be suggested by peer marking only and weak students get lower marks but not as low as their peers would wish which affords a certain protection. From the course leader's viewpoint this offers some reassurance that students will not be overly disadvantaged by potential malicious practice.

Figure 5. Moderation Factors
derived from Different Sources,
5 Teams

Closer examination of the data confirms that a large number of students award marks in higher proportion to themselves than their peers do. However, plots of assessments that included both self-marks and peer marks (the Web-PA scenario) actually provided a better correlation with mentor assessments than when either marks were compared in isolation. These results contrast with the peer assessment of group work study by Orsmond (1996), that poor work was over-marked by peers and good work was under-marked, which could be a function of the small team size in that study (two) and hence, the resulting lack of anonymity.

Figure 5 shows a selection of teams across the spectrum of large and small mark variations and illustrates the broad agreement between all data sources as to who is competent and who is not. Each team has three members that are identified by numbers 1-3. In team E for example, member E1 is clearly the dominant member and member E3 is the weak link and this is confirmed to different degrees by all the respondents. 'Exploiting of peers by unscrupulous students' (Boud et al., 1999) was identified by as a practical problem in peer assessment and this histogram illustrates how the Web-PA system moderates the tendency for peers to award extreme marks. In team A, however, all the indications are that A2 is the strongest member but, interestingly the other two members gave themselves better scores than A2 gave him or herself which further suggests that the weaker members are marking themselves strategically to protect their position, further confirming Boud's findings. This sample data is typical of the whole data set.

Although the present study is also of limited size, the results give some reassurance of the validity of a review system that includes both self and peer-marking. The system clearly identifies who is strong and who is weak and also where a team contains evenly matched members. The authors believe their system implementation has therefore proved to be valid, reliable and transparent as required by the UK Quality Assurance Agency Code of Practice for Assessment (QAA, 2006) and this view has been reinforced by the widespread adoption of Web-PA across the university's faculties, and more recently, the award of an external grant to continue its development with the eventual aim of rolling it out across the UK.

Searching for evidence that teamwork really can deliver academic success

Belbin said, 'Nobody's perfect, but a team can be' (Belbin, 2003). An effective team is not just a band of outstanding players. A team needs a good range of skills; but the key is how the team members work with each other. As previously stated, the basic objective of good teamwork should be to achieve synergy. The authors speculated that synergy might be equated with harmony and that a harmonious team might therefore be expected to outperform a team in conflict with a similar knowledge and skills base. Perhaps the data accumulated from Web-PA would provide the answers?

Year marks from academic transcripts over two completed years were considered as the benchmark for the ability of any given student

and considered alongside peer assessment data from past second year projects. The data was examined for trends which might indicate that harmonious teams (teams with a small variation in peer marking) had exceeded the academic expectations of the team members, however, the results were inconclusive. Overall year marks based on a large number of modules that include examinations and a variety of coursework assignments tend to be consistently lower than marks for the team project activity in any case. Most students like industrially based team projects and they tend to work particularly hard at them. Even allowing for this, it was not difficult to identify evidently harmonious teams that had performed beyond reasonable expectations. On the other hand, it was equally possible to identify many teams that had displayed disharmony by this measure but had still done better than one might expect. A few showed the opposite characteristic. Close inspection shows that some of the largest variations in peer marks occurred in teams of four where three members were approximately equally rated but the fourth was shown to be weak – an indicator, by our metrics, of disharmony. On reflection, it is perhaps not surprising that the overall team project mark was still quite high: the three would take over the task to all intents and purposes, perhaps allocating minor assignments to the weak member and the team of three would continue to perform as a harmonious sub-unit.

The complete lack of identifiable trends led to the conclusion that it is unreasonable to use overall academic performances as an absolute indicator of how well engineering students might tackle open-ended design or research projects. Experience has frequently identified students who are not particularly strong in the examination hall but who excel in a creative team situation. Furthermore, when Web-PA assessment was compared with an equivalent factor calculated from the year marks that related the relative academic strength of each team member to the team average, the conclusion was that project performance is often very different from mainstream academic performance. The evidence appears strong that first class students are not always first class in

respect of team projects and some students who perform quite poorly in examinations can contribute well to team projects. This raises some interesting further questions.

The team projects that form the subject of this study were real world problems set exclusively by partner industrial companies. Industrial tutors were in touch with the teams throughout and were present at review meetings. They also double marked the written reports along with university supervisors. A wide range of industries was involved. It is logical, therefore, to conclude that this project work is about as close as one can get to the kind of task an engineer in industry might tackle. Traditional assessment methods like examinations and laboratory reports do not appear to provide an effective yardstick to measure a student's ability in tackling such a project, so are they still the most appropriate forms of assessment to determine whether our students are adequately prepared for life in industry? Or are real world projects the more accurate determinant of who will perform well upon graduation?

Conclusions

- Self and peer assessments submitted privately by team members against stated criteria broadly agree with the opinions of 'fly-on-the-wall' mentors.
- Within this sample only 1.5% of the results showed widely divergent opinions. This number could be attributed to personality clashes or by mentors misjudging the abilities and efforts of their students.
- Peer-only marking tends to over exaggerate differences between the team members; a combined system of self and peer marking tends to moderate extreme marks and provides a comfort zone for the supervisor, by reducing the number of students with very low marks.
- Individual mark variations that have moderated extremes are an accurate reflection of the rewards for most team situations outside university.
- The research has posed further questions about the validity of traditional examination methods as accurate indicators of the ability of university students to perform in industry.

∎

References

Belbin, M. (2003) *Management Teams - Why they Succeed or Fail.* 2nd edition. Oxford: Butterworth Heinemann.

Boud, D., Cohen, R. & Sampson, J. (1999) Peer learning and assessment. *Assessment and Evaluation in Higher Education,* **24** (4), 413-426.

Crawford, A, R., Tennant, J., Choo, B., S., Wilson, D.A. (Ed). (2003) A Guide to Learning Engineering through Projects, *FDTL(3) report 43/99 PBLE: Project Based Learning in Engineering edited works,* Sept 2003. Available from http://www.pble.ac.uk [accessed 16 May 2007].

Gatfield, T. (1999) Examining student satisfaction with group projects and peer assessment. *Assessment and Evaluation in Higher Education,* **24** (4), 365-377.

Lejk, M. & Wyvill, M., (1996) A survey of methods of deriving individual grades from group assessments. *Assessment and Evaluation in Higher Education,* **21** (3), 267-281.

Orsmond, P., Merry, S. & Reiling, K. (1997) A Study in Self-assessment: Tutor and Students' Perceptions of Performance Criteria. *Assessment and Evaluation in Higher Education,* **22,** 357-369.

Quality Assurance Agency for Higher Education (2006) *Code of practice for the assurance of academic quality and standards in higher education Section 6: Assessment of students,* September 2006. Available from http://www.qaa.ac.uk/academicinfrastructure/codeOfPractice/section6/ [accessed 16 May 2007].

Race P. (2001) *A Briefing on Self, Peer and Group Assessment.* Higher Education Academy, Assessment Series 9. Available from http://www.heacademy.ac.uk/resources.asp?process=full_record&section=generic&id=9 [accessed 16 May 2007].

Tuckman, B, W. (1965) Developmental sequence in small groups. *Psychological Bulletin,* **63,** 384-399. Also reprinted in *Group Facilitation: A Research and Applications Journal,* **3,** Spring 2001, and available as a Word document: http://dennislearningcenter.osu.edu/references/group%20dev%20article.doc [accessed 16 May 2007].

Willmot P. (2003) Introduction of Student Mentors into a Programme of Industry Linked Projects. *Proceedings of International Conference on Engineering Education, ICEE 2003,* Area 7, Paper 1367, Valencia, Spain, July 2003, 1-5 .

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 2004, 1-7.

About the authors

Peter Willmot BSc PhD C.Eng MIED
Principal University Teacher and Director of Undergraduate Studies in the Wolfson School of Mechanical and Manufacturing Engineering, Loughborough University.
School of Mechanical and Manufacturing Engineering, Loughborough University, LE11 3TU, UK.
E-mail: P.Willmot@lboro.ac.uk   Tel: 01509 227555

Adam Crawford BSc MSc PhD
Manager of the UK Engineering Centre for Excellence in Teaching and Learning (engCETL) at Loughborough University.