Assessment & Evaluation in Higher Education
Vol. 33, No. 2, April 2008, 179–190

Relationships between student characteristics and self-, peer and tutor evaluations of oral presentations

A. Mark Langan[a]*, David M. Shuker[b], W. Rod Cullen[a], David Penney[c], Richard F. Preziosi[c] and C. Philip Wheater[a]

[a]*Manchester Metropolitan University, Manchester, UK;* [b]*Edinburgh University, Edinburgh, UK;* [c]*Manchester University, Manchester, UK*

There are many influences on how assessors grade themselves and others. Oral presentations are useful for exploring such factors in peer, self- and tutor marked assessments, being rapidly completed and assessed, commonly used in HE and very difficult to anonymize. This opportunistic study examined the effects of gender and level of attainment on the triangulation of marks awarded to student presenters. Grades generated by peer assessment were associated more strongly with tutor-awarded marks than those from self-assessment. For self-assessment there was a strong effect of gender (female students undervalued their performance compared with tutor grades). Peer assessment produced higher marks than from tutors, perhaps because of the close-knit community developed during residential courses. For tutor marks, the greatest variability was at the lower end of the scale, whereas peer assessors were most variable when marking students who self-evaluated or peer assessed highly. Students awarded a narrower range of marks to peers compared with tutors, but when self-assessing used a larger range. Presentations by students who admitted to little sleep the night before received lower grades from both peers and tutors, but this was not reflected by self-assessments, suggesting they were unaware of their poorer performances. Sessions with fewer talks (four rather than seven) reduced the 'dip' in marks previously observed in the middle of sessions. Findings are discussed in the context of bias in this mode of assessment.

Introduction

Self- and peer evaluation are necessary skills for lifelong learning and professional development (e.g. Sluijsmans et al. 2001) and there is growing support for their inclusion in HE assessment structures (e.g. Hoover & Carroll 1987; Falchikov & Goldfinch 2000). There is considerable literature on the theoretical underpinning of these processes (e.g. Boud 1995a). The current study extends previous work on peer assessment (Langan et al. 2005) by focusing on the relationships between personal traits and self-evaluation.

The term 'self-assessment' has been applied to several distinct activities associated with students making personal judgements on their academic assessment (Boud 1995b). Student ability to self-evaluate has been increasingly important since the implementation of student progress files (e.g. Personal Development Plans) in HE institutions in the United Kingdom. Self-evaluation skills have much in common with those for successful independent study (Clark 1991) and the value of peer assessment is increasingly appreciated (Falchikov & Goldfinch 2000). Inclusion in assessment processes empowers learners and encourages them to assess their peers objectively (Hanrahan & Isaacs 2001). It provides an active learning approach with numerous pedagogical benefits.

*Corresponding author. Email: m.langan@mmu.ac.uk

ISSN 0260-2938 print/ISSN 1469-297X online

DOI: 10.1080/02602930701292498
http://www.informaworld.com

Other potential benefits to learners include developing autonomy, increasing understanding of assessment procedures, and enhancing reflection (e.g. Stefani 1994). Disparity between self-evaluation and tutor (or peer) attainment marks provides insight into learning systems, particularly if key factors can be identified as influencing self-evaluation. For example, interactions between gender and mode of assessment continue to receive attention (e.g. Woodfield et al. 2005).

Incorporation of naive and/or multiple assessors in self- and peer-assessment processes requires advance consideration (Tan 2004) and there is a growing body of literature to guide the design and implementation of appropriate protocols (e.g. Miller 2003). Convergence of marks from different sources is a recurrent issue (e.g. Sim & Sharp 1998) and tutors are often concerned with reliability and validity of marks generated by students. Falchikov & Goldfinch (2000) warn against using agreement between tutors and peers as a measure of validity. Factors including gender, background, engagement with the assessment process and assessment criteria creation can interact with marks originating from student assessors (e.g. Archer 1992).

Considerable attention has been given to gender bias in different modes of assessment. For example, Pirie (2001) argues that approaches such as continuous assessment and 'feminised exams' have recently enhanced female success. Woodfield et al. (2005) note that conscientious female students are suited to coursework, whereas confident, risk-taking males may prefer, but not do better at, unseen examinations.

Self and peer-evaluations of oral presentations are unusual since they are often experienced by student peers and are difficult to anonymize. This is useful for exploring the influences of presenter qualities such as gender and personality type (Falchikov & Magin 1997). Previously, we demonstrated that peer assessors were fairly precise in their marking but tended to over-mark oral presentations compared with tutors (Langan et al. 2005). Potential biases in peer assessment were detected for gender (male markers favoured male speakers), institutional affinity (slight bias towards those from their own university) and peer-group status (anecdotal evidence of 'popular' individuals receiving higher peer marks).

The current study triangulates tutor, peer and self-assessment marks for oral presentations given at the end of two residential courses and investigates quantitatively: (1) the influence of student attributes (gender, university affiliation, participation in creating assessment criteria and amount of sleep prior to the presentation) and presentation attributes (time of day and timing of talk) on the quality of presentation and ability to self- and peer assess; (2) convergence between self-, peer- and tutor-awarded marks; (3) the influence of presentation quality on the variability of marks awarded.

Methods

Data was collected during two residential field courses to southern Spain (2 July–16 2002 and 8–22 July 2003), run jointly between two UK universities. Student numbers varied between courses (n_{2002} = 41; n_{2003} = 19). Full, voluntary participation in the study was given by all staff and students. Student participants willingly volunteered personal information such as the number of hours they slept the night before presenting (a factor of general significance due to student life-styles, and not merely on residential courses). The course format was the same in both years and was based on the format described by Wheater (1989) and summarized by Langan et al. (2005). Most of the students were studying for biological or environmental degrees and this was reflected in the research and teaching specialisms of the tutors (ntutors = 11 on both years: two being female and eight covering both years).

On the final day of each course, students delivered five-minute presentations summarizing their individual research projects, assessed by tutors, a subset of peers and themselves. During the 2002 course a stratified-random selection of 12 students helped to create the assessment criteria

(Langan et al. 2005). The presentations challenge students to distil their intensive two-week projects into concise, clear, informative dialogue aimed at a 'scientific' audience. This is a high-level skill that condenses considerable information into an accessible synthesis. Despite the brevity of presentations (for logistical reasons) we are satisfied that we rigorously assess the quality of the students' work. Students are prepared for the presentations with seminars covering the presentation format, assessment criteria, the use of marking sheets and the concept of peer and self-assessment. Details of the assessment criteria used by all assessors are in Langan et al. (2005). Marks were awarded for presentation (effective/stimulating communication, and clarity of delivery; 40%), content (reference to background, aims, methods, major results and conclusions; 40%) and structure (order, flow and timing; 20%). Markers were given benchmark statements outlining thresholds for excellence and minimal pass levels to assist marking.

Presentations were organized into thematic sessions, chaired by a student. In 2002, sessions comprised six or seven talks; however, a consistent 'dip' in marks awarded to presentations in the middle of sessions (Langan et al. 2005) led to shorter sessions of four presentations being adopted for 2003. In each session the chair and presenters did not peer assess. At the end of all the presentations, students completed a self-assessment form similar to those used for peer assessment. In 2002, this form also asked how many hours they had slept the previous night. For both years, speaker/marker gender, university affiliation, session number and position of the talk in the session were recorded. In 2002, two tutors conducted structured interviews with four random groups of three students covering the peer/self-assessment aspects of the course. Interviews lasted 10–15 minutes and were completed a day after the presentations, allowing time for post-assessment reflection.

Each student's self-assessment mark was analysed together with mean tutor and peer assessment marks. General linear modelling (GLM) was used to characterize self-assessment and explore relationships between self-assessment grades and tutor- and peer-awarded grades. Given the number of variables, two sets of models were generated, one for personal attributes (year, gender, university affiliation, participation in constructing assessment criteria, sleep) and the second for presentation factors (session number, placement within session, and duration of talk). Main effects and interactions were tested by model simplification to yield minimum adequate models, following Crawley (2002). Variation in tutor- and peer-assessment grades for individual students was calculated as coefficients of variation. Pearson's correlation coefficients were calculated to examine trend analyses.

Results

For both years tutor marks for presentation, content and structure were highly significantly positively correlated with the total mark ($r > 0.8$, $p < 0.001$ in all cases), hence further analyses used the latter. A summary of the effects on peer-, self- and tutor-awarded marks is given in Table 1.

Did student attributes relate to tutor grades and the ability to self- and peer assess?

There was no significant difference in self assessment marks between the two years of the study ($F_{1,56} = 0.06$, $p = 0.80$). Over both years, unlike males, female students undervalued their own performance when compared with tutor marks (GLM: $F_{1,57} = 7.07$, $P = 0.01$; Figure 1). Self-assessment was influenced by the university to which a student belonged, i.e. students from one university awarded significantly higher self-assessment marks than those from the other ($F_{1,57} = 4.23$, $p = 0.04$). In 2002, there was no significant effect of student participation in creating the assessment criteria on self-assessment grade ($F_{1,37} = 2.85$, $p = 0.10$); however, effects of gender and university affiliation were detected ($F_{1,38} = 8.76$, $p = 0.006$; $F_{1,38} = 4.66$, $p = 0.04$ respectively;

Table 1. Summary of the effects of learner attributes and presentation structure on self-, peer and tutor marks[1].

		Self-awarded marks (by speakers)	Peer-awarded marks (to speakers)	Tutor-awarded marks
Learner attributes	Gender	Males closer to tutor grades Females differed between universities and were lower than tutor grades	Males marked males higher[2]	NS
	University	Minor institutional differences	Minor institutional differences	Minor institutional differences
	Sleep	NS	Speakers who slept less received lower grades	Speakers who slept less received lower grades
	Participation in creation of assessment criteria	NS. However, effects of participating in assessment criteria generation interacted with the gender and university of the speaker	Participants received slightly lower marks[2]	Participants received slightly lower marks[2]
Structure	Session number	NS	Differences between quality of sessions detected	Differences between quality of sessions detected
	Position of talk in session	NS	Mid-session dip in marks in 2002 (six/seven talks per session) did not occur in 2003 (four talks per session)	Mid-session dip in marks in 2002 (six/seven talks per session) did not occur in 2003 (four talks per session)
	Year	NS	NS	NS

Note: Conclusions are drawn from a range of analyses, where significance was set at $p < 0.05$. NS = no significant effects. [2]For completeness some findings from a previous study are included (Langan et al. 2005).

Figure 2). In addition, there was a significant second-order interaction term in the model (gender*university*participation: $F_{1,33} = 4.50$, $p = 0.04$) suggesting that participation did influence self-assessment, but that its effect differed with regard to both gender and university.

The mean total tutor marks were not significantly different between years ($F_{1,57} = 0.87$, $p = 0.35$; $_{(2002)} = 62.5$, S.E.= 1.32, $n = 41$; $_{(2003)} = 63.8$, S.E.= 2.29, $n = 19$). Over both years, there were no significant effects of gender of student ($F_{1,56} = 0.18$, $p = 0.67$) on tutor grades. For peer assessment, there were significant effects of year ($F_{1,55} = 4.38$, $p = 0.04$) and university affiliation of the student being graded ($F_{1,55} = 6.96$, $p = 0.01$), with grades being slightly higher in 2002. There was also a significant interaction between gender and university affiliation ($F_{1,55} = 4{,}63$, $p = 0.04$), with female students being marked lower than males from the same university in one case, but slightly higher than males in the other.

No significant relationship was found between the amount of sleep reported by a student and his/her self-assessment grade in 2002 ($r_{36} = -0.07$, $p = 0.68$). However, more reported sleep was correlated with higher marks from both tutors ($r_{36} = 0.51$, $p = 0.001$) and peers ($r_{36} = 0.378$, $p = 0.02$). Talks lasted between 95 s and 340 s (= 236 ± 10.3 s) and only three speakers exceeded the 300 s time limit. There was a significant positive relationship between timing accuracy and self-assessed grade ($F_{1,39} = 6.44$, $p = 0.02$), with students giving shorter talks awarding themselves lower marks. In neither year was there an effect of session, or position of talk within a session on self-assessment (all $p > 0.15$). For both years, all interaction terms were non-significant ($p > 0.10$).

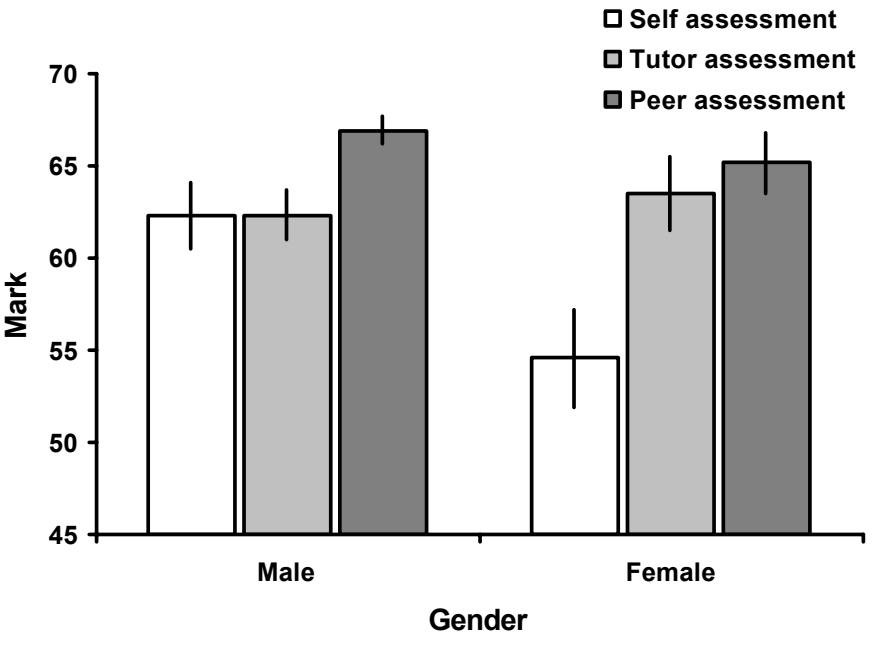

Figure 1. Mean grades of student oral presentations with regard to who assessed the student.
Note: Error bars are standard errors.

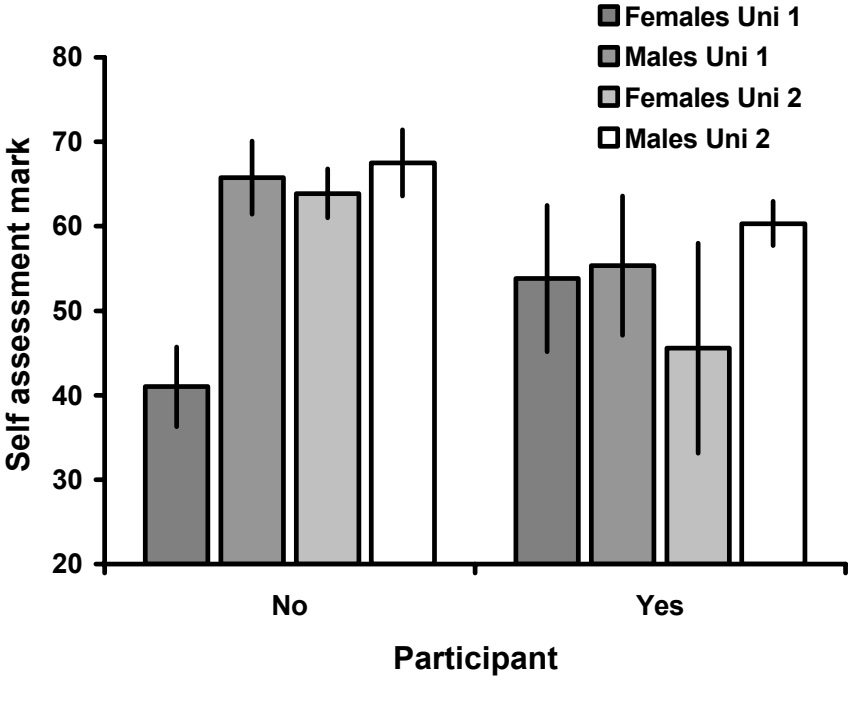

Figure 2. Mean self-assessment grades of student oral presentations with regard to gender, university affiliation and participation in creating assessment criteria carried out in 2002 only.
Note: Error bars are standard errors.

Was convergence achieved between self-, peer- and tutor-awarded marks?

There were significant differences between grades awarded by tutors, students assessing their peers and students assessing themselves (using a mixed-model GLM with graded student as a random factor: $F_{2,118} = 15.28$, $p < 0.0001$, all treatments differed significantly from each other: $p \leq 0.02$ and all interaction terms were removed from the model with $p > 0.2$ in all cases; see Figure 1). In comparison with tutor-awarded grades, students over-marked their peers. However, for self-assessment there was an overall effect of females under-marking themselves.

Tutor marks were highly significantly positively correlated with peer assessment marks (linear regression: $\beta = 1.02 \pm 0.11$, $R^2 = 0.59$, $n = 60$, $p < 0.0001$; Figure 3a) but less so with self-assessment marks ($\beta = 0.20 \pm 0.09$, $R^2 = 0.10$, $n = 60$, $p = 0.01$; Figure 3b). This relationship was independent of gender, year and university affiliation (GLM: all $p > 0.1$). Although Figure 3a suggests that students gave slightly higher marks to weaker students and slightly lower marks to better students compared with tutors, the slope of the regression line for tutor against peer assessment was not significantly different from 1 ($p = 0.89$), even after removal of the outlying low-graded student ($p = 0.17$). Figure 3a also illustrates that, barring one extremely low outlier, peers awarded grades in a range approximately half as wide as tutors. Self-assessment, however, produced a broader range of marks than tutor assessment (Figure 3b) and most underestimated their performance compared with tutor grades. Self- and peer-assessment grades were highly significantly positively correlated ($\beta = 0.86 \pm 0.22$, $R^2 = 0.21$, $p = 0.0002$).

Did variability of marks change with the standard of presentation?

Variation amongst tutor marks (calculated as coefficients of variation for 2002 assessments) was highly significantly negatively correlated with mean tutor grade ($r_{39} = -0.51$, $p < 0.001$), i.e. tutor marks were more variable for lower achieving students (Figure 4a). There was no correlation between variation in tutor grades and self-assessed grades ($r_{39} = 0.05$, $p = 0.74$), so how students felt they did was not correlated with the level of disagreement between tutors. However, variation in peer assessment grades was significantly positively correlated with self-assessment, thus there was more disagreement amongst peer assessors for students who graded themselves more highly than for students who rated themselves poorly ($r_{39} = 0.46$, $p = 0.002$; Figure 4b). Variation in peer assessment was not related to tutor assessment ($r_{39} = 0.25$, $p = 0.12$) although it was significantly positively correlated with peer-assessment grades implying that there was more disagreement between peer assessors over high marks than over low ones ($r_{39} = 0.48$, $p = 0.001$).

Student feedback

Post-assessment interviews revealed that students felt they were more confident in assessing peers than themselves. Some students felt that practice in peer assessment before the presentations (e.g. using a tutor's presentation as an open example) would have increased confidence. Several interviewees noted that their confidence to peer assess increased but ability to concentrate dwindled as the day progressed. A third of the interviewees suggested that they would have concentrated better had they not been peer-marking. This was at odds with tutors' experience that inclusion of peer and self-assessment had increased student engagement in sessions. Benchmark statements were considered useful/essential for discriminating between the performances of speakers, although marking was more difficult when the assessor knew less about the subject area. Five of the interviewees raised issues of feeling uncomfortable about being asked by peers about the marks they awarded and the increased difficulty of self-assessment (compared with peer assessment). All interviewees deemed both peer and self-assessment useful and thought-provoking experiences. One person suggested the addition of at least a median descriptor to benchmark statements.

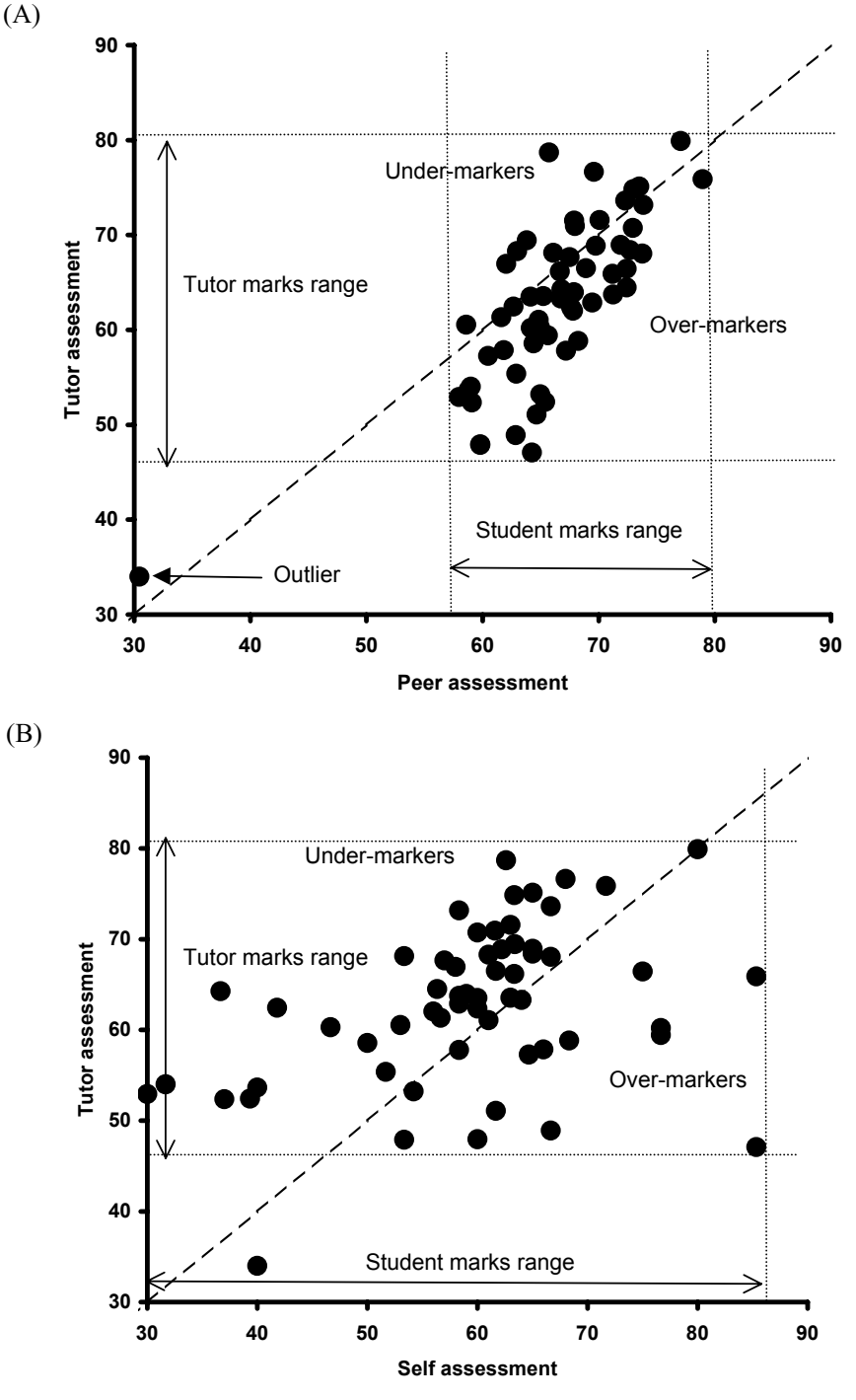

Figure 3. Relationship between tutor assessment and (A) peer assessment, (B) self-assessment.
Note: Dashed lines represent a 1:1 relationship. Dotted lines demonstrate the limits of the majority of marks awarded on the basis of who assessed the student (one extremely low outlier has been omitted from this).

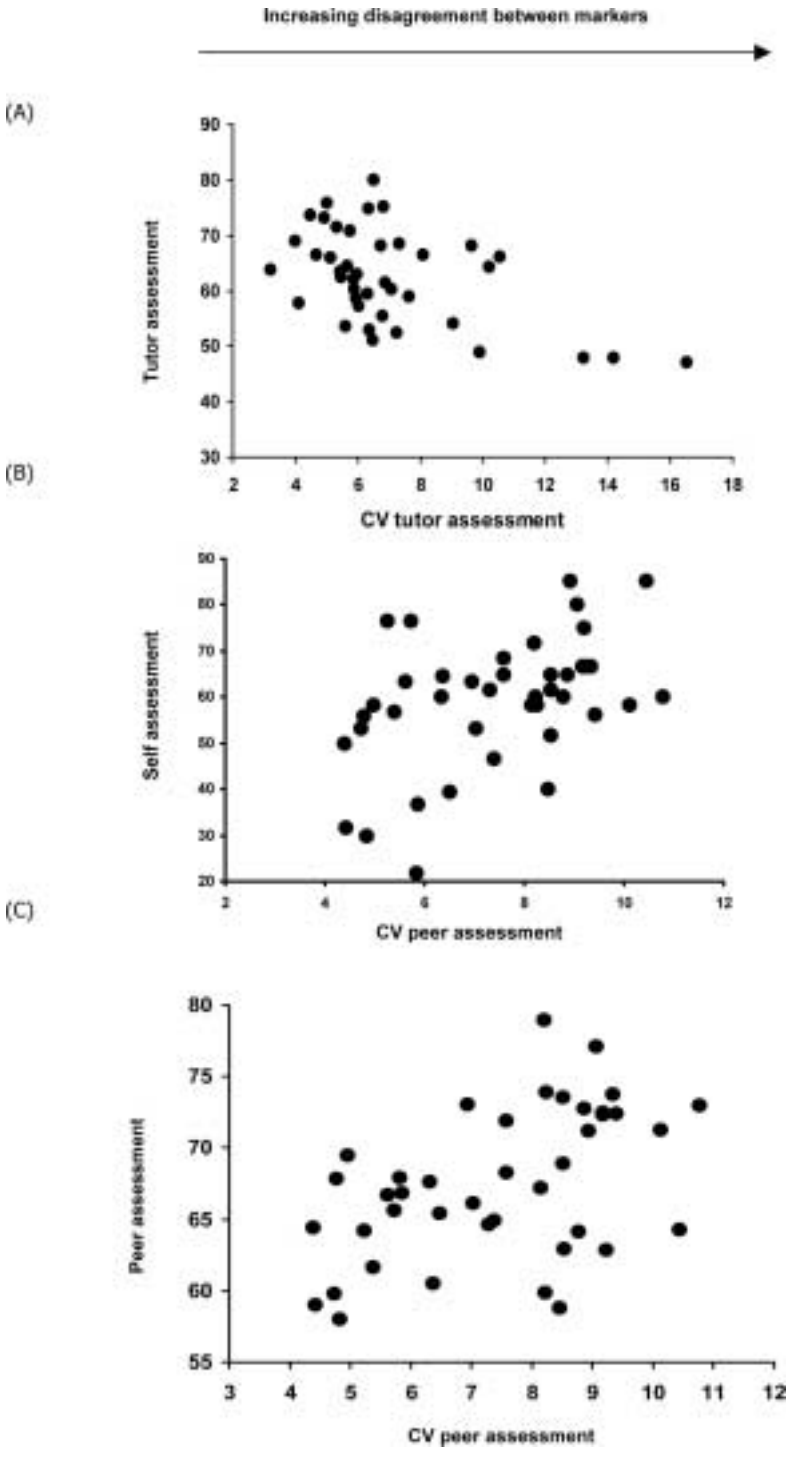

Figure 4. Relationship between the coefficient of variation (CV) and marks for (a) tutor marks, (b) self-assessment grade and peer-assessment variation and (c) peer marks.
Note: Increasing CV values indicate increasing disagreement between markers.

Discussion

The ability to self-assess is crucial for lifelong learning and is increasingly important in a flexible employment market (Orsmond 2004). As in previous work (e.g. Cowan 1988), feedback during interviews suggested that participation in self- and peer assessment provoked thought about the assessment process. Peer and, in particular, self-assessment were a challenge to many students, reflected in part by the high variability in self-assessment marks and their lack of congruence with tutor and peer marks. The nature and number of criteria used in assessment influence the marks generated (see Miller 2003). Feedback in interviews suggested that the threshold descriptors provided were useful, usable and appropriate. However, modifications to a Likert scale are under consideration as this has been shown to enhance student discrimination (MacAlpine 1999).

Understanding the processes of self- and peer assessment requires an appreciation of students' perceptions of themselves and others. Such understanding should enhance future implementations. Factors for consideration include gender, social context (e.g. university affiliation), state of mind (e.g. tiredness), experience of assessment (e.g. inclusion in the creation of assessment criteria) and assessment structure (e.g. position of the presentation within and between sessions). Our findings are consistent with those in Langan et al. (2005), showing that both gender and university affiliation influenced how students assessed. Previously we showed that during peer assessment of presentations, males marked other males higher than females (Langan et al. 2005). In the current study, females from one university undervalued their own performance, whilst males were generally over-valued by peers in comparison with tutor grades.

Interestingly, females involved with the creation of the assessment criteria did not undervalue their work as much (based on a small sample of only three women). These findings add to the equivocal findings of gender effects in the literature, which vary particularly with (amongst other factors) mode of assessment. Oral presentations have been viewed as more 'male-orientated' (Falchikov & Magin 1997), which may at least partly explain male and female differences in this study. Gender differences in the workplace and in education varying according to the context of the observations. For example, students have evaluated male teachers more highly than female counterparts (Feldman 1992); bias in peer-assessed group work has been shown towards females (Falchikov & Magin 1997); and preference for same-sex subjects during assessment has been noted (Eagly & Carly 1981). There are also many examples of studies that did not detect gender differences (e.g. Mattheos et al. 2004).

Unlike peer grades, self-assessment was not strongly associated with tutor grades (due in part to females undervaluing their performances). It is common that self-ratings are poorly correlated with other performance measures, such as tutor-awarded grades (Eva et al. 2004). When compared with marks awarded by tutors, students awarded themselves a broader range of marks (30–85%) but awarded their peers a much narrower range (the majority from 57–72%). This suggests a lack of confidence or ability to discriminate peers, an effect often found with peer and self-assessment (see Miller 2003).

Lower achieving individuals tended to over-mark themselves compared with high achievers, supporting findings that people unskilled in social and cognitive domains make erroneous, often inflated, judgements about themselves (Kruger & Dunning 1999). High self-assessment marks may reflect high levels of confidence (i.e. to some extent independent of the quality of presentation) or poor understanding of academic level in relation to the requirements of the assessment. It is worth considering that the nature of benefits of self-evaluation may be complex if learners are being disciplined by the process more than encouraged to be autonomous in their thinking (Tan 2004) or lack the cognitive abilities to engage with, or reflect fully upon, these experiences at a level appropriate to the task (Kruger & Dunning 1999). There is a need to develop exercises like those in the current study to continue to utilize and apply the feedback in a series of exercises

(e.g. Ellery & Sutherland 2004) to permit evaluation as to whether engagement with these processes leads to increased abilities to peer and self-assess. Ultimately, Cowan (1988) argued that the learning benefits of self-assessment outweigh the differences between student- and tutor-derived grades. It is noteworthy that tutors also disagreed, in this case most markedly over marks for lower attainment students, whilst peer assessors tended to disagree more over high-quality speakers indicated by both self- and peer assessment.

In consideration of factors associated with the presentations, self-assessment was not related to the position of a talk within or between sessions. In 2002 we observed a mid-session dip in marks (Langan et al. 2005). This was not seen in shorter sessions run in 2003. Loss of concentration and low activity between 20 and 40 minutes have been widely reported elsewhere (e.g. Penner 1984), consequently we recommend shorter sessions. In addition, individuals who gave shorter talks gave themselves a poorer mark, which also reflected in peer and tutor assessment. This may result from the amount of information disseminated and, perhaps, the behaviour of speakers finishing early (for example ending with 'Well … errr …. that's it').

Lack of sleep is a relevant aspect of residential courses and unusual to document. The number of hours' slept before the assessment day correlated positively with the mark attained from both tutors and peers, but may be associated with many confounding variables (e.g. personality type, age, commitment to the course). It was interesting that with self-assessment this relationship was not detected, indicating that sleep-deprived students lacked the ability to evaluate themselves, although this interpretation is subject to the same confounding variables.

The level of convergence among self-, peer- and tutor-generated marks in our study was complex and explained in part by gender effects. Overall, self-assessment grades were lower than either tutor grades or peer grades. This was primarily due to underestimation by females of their own performances, since males self-assessed with high accuracy when compared with tutor-awarded marks. Self-assessments did correlate significantly with tutor grades but not as highly as peer assessment. Falchikov & Goldfinch (2000) noted that self-assessment is intended as a private activity and probably involves little knowledge of the work or performance of others. Strong association but low accuracy of self-assessments with tutor grades may reflect a lack of experience of marking and a need for calibration of self-assessment through training (as suggested for peer-awarded grades in Langan et al. 2005) and coaching (Schelfout et al. 2004). Such training would be the appropriate stage to raise awareness of factors contributing to bias in marking (e.g. gender, background).

The association between tutor- and student-awarded grades (self or peer) can be influenced by characteristics of the students/courses involved, and previous authors have found strong associations between self- and tutor assessments (e.g. Stefani 1994). Falchikov & Boud (1989) suggested that advanced science courses seemed to produce more accurate self-assessment. The current students were 'fairly advanced' (at the end of their second-year undergraduate stages). However, there is opportunity to improve self-assessment abilities by training students. The residential format may interact with the psychology of student evaluation of friends, colleagues and themselves, which may explain some of the patterns seen in this study. The intense residential format facilitates cohesive peer groups, but can be a stressful personal experience in certain contexts, but may have led to harsher self-analysis and generosity towards peers. Increasing the number of student assessors can improve correspondence between tutor- and student-generated marks (e.g. Magin 1993). However, the current study did not detect differences in student marking accuracy between years despite over double the number of students taking part in 2003 (41) compared with 2002 (19) agreeing with Falchikov & Goldfinch (2000) whose study comprised a comprehensive meta-analysis of peer-awarded grades.

In summary, female students undervalued their own performance in oral presentations over two years, and this effect was related to university affiliation. The opposite occurred with peer

assessment where students over-marked colleagues from the same institution compared with tutor grades. Greater variability in marks was observed for the lower end of the scale for tutors, and for peer assessors of students who self-evaluated or were peer assessed highly. Modification of the structure of the presentations to shorter sessions removed the 'dip' in marks previously awarded in the middle of sessions. The amount of sleep students had prior to the presentation day was strongly correlated to the final grade, but we could not distinguish confounding variables to elucidate the major causes of this relationship. Overall, tutor assessments were most closely associated with peer assessment, rather than self-awarded grades. However, peer assessment was constrained by a narrow mark range, unlike self-assessments.

Acknowledgements

The authors would like to thank Robin Baker, Carl Ashcroft, Gordon Bennett, Jen Boyle, Matt Cobb, Ben Haines, Debra Hamilton, Les Lockey, Johan Oldekop, Emma Shaw and all the students who have made this work possible. DMS thanks Stu West for his support in undertaking this collaborative project. Thanks are also offered anonymous referees for their comments on the manuscript.

Notes on contributors

Mark Langan is a senior learning and teaching fellow who researches assessment strategies, learning styles and the learning experiences of international students. For his subject research he explores the impact of environmental stressors on invertebrates.

Dave Shuker is a post-doctoral research fellow in the School of Biological Sciences at the University of Edinburgh, studying evolutionary ecology and teaching field biology and data analysis.

Rod Cullen specializes in online and distance learning, formerly at the University of Manchester and is now employed in the Learning and Teaching Unit at Manchester Metropolitan University.

David Penney is a post-doctoral researcher of fossil and extant spiders in the Department of Earth, Atmospheric and Environmental Sciences at the University of Manchester and teaches field biology.

Richard Preziosi is a lecturer and researcher of population genetics in the School of Biological Sciences at the University of Manchester with interests in the design of field-based courses.

Phil Wheater is a principal lecturer researching human impacts on ecological systems and has taught field ecology for over 30 years, first at the University of Manchester and now at the Manchester Metropolitan University.

References

Archer, J. 1992. Sex bias in evaluations at college and work. *The Psychologist: Bulletin of the British Psychological Society* 5, no. 5: 200–04.

Boud, D. 1995a. *Enhancing learning through self- assessment.* London: Kogan Page.

———. 1995b. Developing a typology for learner self assessment practices. *Research and Development in Higher Education* 18: 130–35.

Clark, R. 1991. Student opinion of flexible teaching and learning in higher education. In *Flexible learning in higher education,* eds. W. Wade, K. Hodgkinson, A. Smith, and J. Arnfield, 136–50. London: Kogan Page.

Cowan, J. 1988. Struggling with student self-assessment. In *Developing student autonomy in learning,* ed. D.J. Boud 192–210. London: Kogan Page.

Crawley, M.J. 2002. Statistical *computing: an introduction to data analysis using S-Plus.* Chichester: Wiley.

Eagly, A.H., and L.L. Carly. 1981. Sex of researchers and sex-typed communications as determinants of sex differences in influenceability: a meta-analysis of social studies. *Psychological Bulletin* 90: 1–20.

Ellery, K., and L. Sutherland. 2004. Involving students in the assessment process. *Perspectives in Education* 22, no. 1: 99–109.

Elwood, J. 1999. Gender, achievement and the 'Gold Standard': differential performance in the GCE A Level examination. *Curriculum Journal* 10, no. 2: 189–208.

Eva, K.W., J.P.W. Cunnington, H.I. Reiter, D.R. Keane, and G.R. Norman. 2004. How can I know what I don't know? Poor self-assessment in a well-defined domain. *Advances in Health and Sciences Education* 9: 211–24.

Falchikov, N., and D. Boud. 1989. Student self assessment in higher education: a meta-analysis. *Review of Educational Research* 59: 395–430.

Falchikov, N., and J. Goldfinch. 2000. Student peer assessment in higher education: a meta-analysis comparing peer and teacher marks. *Review of Educational Research* 70: 287–322.

Falchikov, N., and D. Magin. 1997. Detecting gender bias in peer marking of students' group process work. *Assessment and Evaluation in Higher Education* 22, no. 4: 385–96.

Feldman, K.A. 1992. College students' views of male and female college teachers, Part I: Evidence from social laboratory experiments. *Research in Higher Education* 33, no. 3: 317–75.

Hanrahan, S.J., and G. Isaacs. 2001. Assessing self- and peer-assessment: the students' views. *Higher Education Research and Development* 20, no. 1: 53–70.

Hoover, N.L., and R.G. Carroll. 1987. Self-assessment of classroom instruction: an effective approach to in service education. *Teaching and Teacher Education* 3: 179–91.

Kruger, J., and D. Dunning. 1999. Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *American Psychological Association* 77: 1121–34.

Langan, A.M., C.P. Wheater, E.M. Shaw, B.J. Haines, W.R. Cullen, J. Boyle, D. Penney, J. Oldekop, C. Ashcroft, L. Lockey, and R.F. Preziosi. 2005. Peer assessment of oral presentations: effects of student gender, university affiliation and participation in the development of assessment criteria. *Assessment and Evaluation in Higher Education* 30: 19–34.

MacAlpine, J.M. 1999. Improving and encouraging peer assessment of student presentations. *Assessment and Evaluation in Higher Education* 24: 15–25.

Magin, D. 1993. Should student peer ratings be used as part of summative assessment? *Research and Development in Higher Education* 16: 537–42.

Mattheos, N., A. Nattestad, E. Falk-Nilsson, and R. Attstrom. 2004. The interactive examination: assessing students' self-assessment ability. *Medical Education* 38: 378–89.

Miller, P.J. 2003. The effect of scoring criteria specificity on peer and self-assessment. *Assessment and Evaluation in Higher Education* 28: 383–94.

Orsmond, P. 2004. *Self- and peer- assessment: guidance on practice in the biosciences,* Teaching Bioscience Enhancing Learning Series, Centre for Bioscience Leeds: Higher Education Academy.

Penner, J.G. 1984. *Why many college teachers cannot lecture: how to avoid communication breakdown in the classroom.* Springfield, IL: Charles C. Thomas.

Pirie, M. 2001. How exams are fixed in favour of girls. *The Spectator* 20 January: 12–13.

Schelfout, W., F. Dochy and S. Janssens. 2004. The use of self, peer and teacher assessment as a feedback system in a learning environment aimed at fostering skills of cooperation in an entrepreneurial context. *Assessment & Evaluation in Higher Education* 29: 177–201.

Sim, J., and K. Sharp. 1998. A critical appraisal of the role of triangulation in nursing research. *International Journal of Nursing Studies* 35: 23–31.

Sluijsmans, D.M., G. Moerkerke, J.J. Merrienboer, and F.J. Dochy. 2001. Peer assessment in problem-based learning. *Studies in Educational Evaluation* 27: 153–73.

Stefani, L. 1994. Peer, self and tutor assessment: relative abilities. *Studies in Higher Education* 19: 69–75.

Tan, K.H. 2004. Does student self assessment empower or discipline students? *Assessment and Evaluation in Higher Education* 29: 651–62.

Wheater, C.P. 1989. A comparison of two formats for terrestrial behavioural ecology field courses. *Journal of Biological Education* 23, no. 3: 223–31.

Woodfield, R., S. Earl-Novell, and L. Solomon. 2005. Gender and mode of assessment: should we assume female students are better suited to coursework and males to unseen examinations? *Assessment and Evaluation in Higher Education* 30: 35–50.