Review of Educational Research
Fall 2000, Vol. 70, No. 3, pp. 287-322

Student Peer Assessment in Higher Education: A Meta-Analysis Comparing Peer and Teacher Marks

Nancy Falchikov
Judy Goldfinch
Napier University

Forty-eight quantitative peer assessment studies comparing peer and teacher marks were subjected to meta-analysis. Peer assessments were found to resemble more closely teacher assessments when global judgements based on well understood criteria are used rather than when marking involves assessing several individual dimensions. Similarly, peer assessments better resemble faculty assessments when academic products and processes, rather than professional practice, are being rated. Studies with high design quality appear to be associated with more valid peer assessments than those which have poor experimental design. Hypotheses concerning the greater validity of peer assessments in advanced rather than beginner courses and in science and engineering rather than in other discipline areas were not supported. In addition, multiple ratings were not found to be better than ratings by singletons. The study pointed to differences between self and peer assessments, which are explored briefly. Results are discussed and fruitful areas for further research in peer assessment are suggested.

Student involvement in assessment appears to have been increasing in recent years, if a cursory review of the literature in higher education research is to be believed. This increase appears across the spectrum of discipline areas including science and engineering, arts and humanities, mathematics and education, and social sciences and business studies, and across a very wide range of student experiences from pre-course to advanced stages and even in post-course professional practice. Furthermore, the burgeoning research literature on peer assessment suggests that student involvement is a world-wide phenomenon.

Student Self and Peer Assessment

Student involvement in assessment typically takes the form of peer assessment or self assessment. In both of these activities, students are engaging with criteria and standards, and applying them to make judgements. In self assessment, students judge their own work, while in peer assessment they judge the work of their peers. Peer assessment is grounded in philosophies of active learning (e.g., Piaget, 1971) and androgogy (Cross, 1981), and may also be seen as being a manifestation of social constructionism (e.g., Vygotsky, 1962), as it often involves the joint construction of knowledge through discourse. Peer as-

sessment activities have been found to promote learning (e.g., Boud, 1988; Falchikov, 1986), and it is this aspect which commonly forms the rationale for introducing peer assessment into courses. An important educational function of peer assessment is the provision of detailed peer feedback (Falchikov, 1994, 1995). Topping (1998) includes a useful case study of peer assessment which illustrates many aspects of the technique.

Issues of Reliability and Validity

Fears of teachers about the lack of reliability or validity of peer assessment may act to restrict its use and, thus, deprive many students of its learning benefits. Does the present study aim to investigate reliability or validity of student marking? If our primary concern is the agreement between peer ratings, then we could be said to be examining reliability. If, however, we are validating students' ratings against those of teachers as a standard, then it can be argued that our concern is with validity. Work in the area of marking or grading is fraught with difficulty, as teacher marking has, itself, been found to be problematic (e.g., Falchikov & Magin, 1997; Guilford, 1965; Newstead & Dennis, 1994). In fact, Guilford argued that marks may be neither very reliable nor very valid indicators of achievement, and Marcoulides and Simkin (1991) argued that even when there is a reasonable degree of agreement between raters, "consistent grades are not necessarily 'fair' grades" (p. 82). For example, Newstead and Dennis (1990) argued that several different kinds of bias in marking might operate.

However, the main concern of many teachers is the degree of agreement between their marks and those awarded by their students. Thus, although many of the studies contributing to the meta-analysis claim to be reporting data that relate to reliability of ratings, we conceive of the present study as an investigation of the validity of peer marking.

Falchikov and Boud's (1989) Meta-Analytic Study of Student Self Assessment Studies

Falchikov and Boud (1989) subjected 57 quantitative self assessment studies which compared self and teacher marks to a meta-analysis. Important factors with regard to the closeness of correspondence between self and teacher marks were found to include the quality of design of the study, the level of the course of which the assessment was a part, and the subject area in which the assessment took place. Better designed studies were associated with closer correspondence between student and teacher than poorly designed ones, and students on advanced courses appeared to be the more "accurate" assessors than those on introductory courses. Studies within the broad area of science seemed to produce more accurate self assessment generally than those from other discipline areas.

The present study may be seen as a companion piece to Falchikov and Boud (1989). Both focus on the marking aspect of student involvement in assessment and both state a belief that self and peer assessment involves a great deal more than this and that the primary benefit of involving students in assessment resides in the improvement to learning which can result. The two studies share a similar structure, but the present study uses more recent meta-analytic tech-

niques than the earlier one and attempts a preliminary investigation of interactions between variables.

Topping's (1998) Review of Peer Assessment Studies

A recent qualitative review of peer assessment studies by Topping (1998) that focused primarily on the mechanisms and benefits of peer assessment located some studies which compared teacher and peer marks. Topping's review provides a useful starting point to an integrative study of peer assessment in higher education. However, the present study differs from Topping's in several important aspects:

1. Topping's account of peer assessment studies in higher education is qualitative and mainly descriptive, and lacks the means to investigate variations in outcomes. The present study is more narrowly focused on faculty-student marks comparisons and attempts a quantitative analysis of the effects of some key variables in this aspect of peer assessment.
2. Topping's review located only 31 studies which compared teacher and peer marks, compared with 48 in the present study. The present study contains 30 studies not included in Topping's. Of those studies not included in the present corpus, eight were comparisons of *self* and peer assessments, and four were excluded as they did not contain sufficient statistical data or raw data to enable their inclusion. The characteristics of the two quantitative sets of peer assessment studies may differ. For example, Topping claims that "one assessor to one assessee was the modal constellation" (p. 252). Of the 48 quantitative studies examined in the present study, only nine conformed to this pattern.
3. Topping examined the reliabilities of his 31 quantitative studies, basing his conclusions on reported statistics and researcher interpretations.
4. Topping's main parameters of variation have no explanatory power in terms of either the reliability of peer assessment or the perceived learning benefits to students. For example, investigation of the "curriculum area / subject" parameter informs the reader that peer assessment occurs across a wide range of subject areas. It gives no indication of any discipline or subject differences. Topping's typology differentiates between individual assessors, pairs and groups, but his review does not investigate which type of grouping is likely to be most successful in terms of peer assessment reliability. The present study begins to explore the relative importance of these and other variables.

Researcher Interpretation Bias

An important limitation of Topping's (1998) review is that it relies entirely on the interpretations of researchers who may interpret their findings in ways that are not shared by others. For example, Hughes & Large (1993) reported that faculty and peer means and standard deviations were close, but, when calculated, the effect size indicated only a moderate correspondence between the groups. Similarly, Jacobs et al. (1975), investigating differences in peer assessment expertise between groups with differing ability levels, concluded that "student evaluation should not be based on peer ratings by poor students" (p. 540). However, on closer inspection of the correlations between faculty and

student peers, while Jacobs et al.'s conclusion holds for the lowest Grade Point Average (GPA) group, the correlation between faculty marks and those in the "below average" GPA group were very similar to those comparing faculty and the highest GPA group. Furthermore, marks awarded by average-ability students corresponded less well with faculty marks than those in the below average group. Thus, it appears that a somewhat complex set of results have been over simplified by the authors. Authors sometimes over generalize their results. For example, Korman & Stubblefield (1971) claimed that "the best predictors of future internship success turn out to be each student's peers . . . . [T]he peer group evidenced much higher correlations" than other groups involved. On inspection, while the researchers are reporting their results accurately, the value of r associated with the peer group was a somewhat modest 0.14. Thus, there are several problems inherent in accepting the interpretations of researchers.

However, "trustworthy accounts of past research are a necessary condition for orderly knowledge building" (Cooper, 1998, p. 1) and some kind of research synthesis is needed. A meta-analysis is a technique which provides both.

The Meta-Analytic Technique: Its Advantages and Limitations

Qualitative research syntheses are subject to experimenter and reviewer bias. Researchers may introduce bias into their papers when they interpret their findings, as illustrated above. Small effects may be interpreted as large; unwanted or unexpected outcomes may be played down or ignored. Reviewers, too, may introduce bias.

For the present study, a meta-analysis of quantitative peer assessment studies was chosen in order to investigate teacher-student peer agreement in marking. Meta-analysis allows the evidence from different studies to be combined so that individual studies become data points in a large population of studies. In meta-analysis, data on which inferences are drawn are public and open to debate. Another important feature of meta-analysis is that it does not prejudge research findings in terms of the quality of research. Cooper (1998) summarizes the debate on whether or not to exclude some studies a priori on the basis of poor methodology. Some researchers (e.g., Eysenck, 1978) have argued for exclusion of poor studies on the grounds that only better designed experiments can lead to better understanding of issues. Others (e.g., Glass and Smith, 1978) argue for inclusion, reasoning that a priori quality judgements are likely to vary from judge to judge and that poor design features can, in any case, cancel each other out. Cooper himself advocates the inclusion position believing it to be more consistent with a rigorous approach to research synthesis.

However, meta-analysis has its limitations. These include the effects of publication bias, heterogeneity of studies and problems of combining studies with very different sample sizes. Each of these is considered in the analyses presented here. Begg (1994) argues that sampling methods may go some way towards correcting publication bias. If, as is the case here, the corpus of studies includes unpublished work, work in progress and conference presentations where the peer review process may not be as rigorous as in the case of published studies, then the effects of publication bias may be reduced. However, despite one's best efforts, one can never be sure that the search for studies has been exhaustive, as has been illustrated above. Heterogeneity of studies and publication bias were examined in the present meta-analysis and results reported below.

A Meta-Analysis of Quantitative Peer Assessment Studies

Method

Selection of Studies for Inclusion

Peer assessment studies for the present analysis were found by searching the following databases: Bath Information Data Service (BIDS), Educational Resources Information Center (ERIC), MEDLINE, Psychinfo, Socinfo, FirstSearch. Keywords used were *peer, assessment / marking / grading / evaluation, student, higher education.* The search was limited to work in the English language. Bibliographies and review articles were inspected and citations were followed up. Some authors were contacted directly and unpublished work obtained. While direct contact added a few studies to the corpus, no additional information sought regarding published studies was forthcoming. In total, in excess of 100 studies were located. Forty-eight of these were review papers and qualitative accounts of peer assessment in higher education. Forty-eight were quantitative studies that included comparisons of numerical marks or grades awarded by peers and faculty. These spanned the period 1959 to 1999. In each of these, the peer score was typically a mean value derived from several individual peer assessments. The remaining studies involved peer assessment in non–higher-education settings. The selection criteria were that each study must be situated within higher education and that it must contain correlation coefficients or proportions of cases where peer marks were deemed to be in agreement with faculty grades or statistical data to enable calculation of effect sizes. No quality filter was applied at this point.

Coding quantitative peer assessment study characteristics

For each study, the variables that might influence the outcomes were noted (independent variables), as were the findings (dependent variables). Classification of each study was made under the following headings:

Independent variables.

- study identifiers (name of researchers and date)
- population characteristics (number of participants overall, gender, level of students)
- what is assessed
- the level of the module or course (e.g., introductory or advanced)
- how the assessment is carried out and the nature of the criteria used (if known)
- the design quality
- number of peers, and number of faculty involved in assessments

Studies used one of two terms to describe the assessments. Some reported the awarding of marks while others involved grading. While "grading" can indicate awarding students a mark out of 10 or rating them within a band of marks (e.g., the range 70% to 85%), it may also refer to other labels (e.g., A, B). In some cases, it may also be used to indicate the award of marks (often percentages) as in "marking." Thus, in order to minimize confusion, we use the terms "marks" and "marking" throughout this paper to indicate numerical assessment.

Design quality

Bangert-Drowns, Wells-Parker, and Chevillard's (1997) key features of study quality assessment were used in the present study. Criteria for judgements of study quality were explicit; the procedure for determining quality was systematic; and the criteria used have face validity and reflect consensual opinions in the research community (e.g., Falchikov and Boud, 1989). A multivariate strategy with summative scores for methodological quality was adopted. It was deemed important that any high quality study should report enough information to enable replication. This requirement informed the choice of criteria for determining study quality which were

- Inadequate reporting of population characteristics (e.g., age, gender)
- Inadequate reporting of other study characteristics (e.g., type or level of course, number of peer markers per assessment)
- Very small sample size and inappropriate generalizations
- Inappropriate tasks required of students (e.g., prediction of grades or marks, assessing different aspects of performance/ assessing in different ways from teachers)
- Students not provided with criteria or structure; global rating required
- Inappropriate criteria used (e.g., effort)
- Inadequate procedural information (study not replicable from the information given)

One mark was given for each of the study faults. Those studies where information was missing or inadequate in at least four of the areas above were rated as having (or reporting) a poor experimental design (n = 11). All other studies were rated as high quality studies (n = 45).

Dependent variable. The dependent variable was the outcome of each study. A value of a common metric was either supplied directly by the researcher or calculated. Common metrics used were the effect size (d), correlation coefficient (r) or percentage agreement (%).

Table 1 provides a summary of the characteristics and common metrics of the quantitative studies included in the meta-analysis. As in Falchikov & Boud's (1989) meta-analytic study of quantitative self assessment studies, each highlighted characteristic may be regarded as the basis for an hypothesis concerning the relationships between independent and dependent variables.

Quantifying the Experimental Effect: Calculation of Common Metrics

Integration of quantitative results of many studies requires a common statistic. Our corpus gave rise to three: effect sizes, correlation coeffiecients, and proportions.

(a) Effect size calculation

The formula provided by Cooper (1998) was used to calculate effect sizes in cases where means and standard deviations were supplied:

$$d = \frac{\text{(E group mean) - (C group mean)}}{\frac{\text{(E group sd + C group sd)}}{2}}$$

where C = Control and E = Experimental.

It should be noted that, as peer assessment studies are not "true" experiments and have no experimental or control groups in the generally understood sense,

faculty markers were designated the control (C) group and peer markers the experimental group (E). The E group mean is plotted as a *z* score within the C group distribution (cf. Falchikov & Boud, 1989). The greater the distance between means, the greater the absolute difference in performance of the two groups. The "effect size" is a standardized index of deviation in a situation where minimal deviation is required, so small effect sizes are sought, in contrast to the more usual application where larger effect sizes are desired. In the present case, the smaller the absolute effect size the greater the resemblance between student peer markers and faculty markers. Positive d values indicate that peers tend to be more generous in their marking than faculty (referred to as "over marking"), and a negative *d* value indicates the opposite ("under marking").

(b) Correlation coefficients.

Although it is technically possible to convert correlation coefficients *(r)* into effect sizes *(d)* (Glass, McGaw, & Smith, 1981; Hunter, Schmidt, & Jackson, 1982), given the characteristic nature of peer assessment studies, it was deemed more appropriate to regard the correlation coefficient itself as the dependent variable (cf. Hembree, 1988; Falchikov & Boud, 1989). In quantitative self and peer assessment studies, the comparison is between marks awarded to the same process or product by faculty and students, while other studies report correlations between different treatments or methods. The correlation coefficient *(r)* resulting from the second of these types of studies may be transformable into an effect size *(d)* while those from the former may not (Falchikov & Boud, 1989). In addition, Hedges and Olkin (1985) argued that the correlation coefficient is "a natural candidate as an index of effect magnitude suitable for cumulation across studies" (p. 223). Thus, in the present study, correlation coefficients were taken as a direct measure of the dependent variable.

(c) Proportions

Some studies in the peer assessment corpus reported the proportion of comparisons that indicated "agreement" between faculty and students rather than other statistics. Although proportions, like correlations, may be transformed into effect sizes (Glass et al., 1981), this procedure was again deemed inappropriate in the context of peer assessments. First, when we are comparing teacher and student marks, the teacher mark is taken as a standard against which to compare peer assessed marks. Thus, the proportion of teacher "successes" must always be 1.00, and the Bayesian estimate for extreme cases must be used in every comparison. It was, therefore, decided to make direct comparisons between percentages of faculty-student agreement. However, studies reporting proportions of such agreements employed different definitions of "agreement," which varied from the draconian identical ratings of Gray (1987) or Orsmond, Merry, and Reiling (1996) to Lennon's (1995) less than ten marks difference. Others, such as Kwan and Leung (1996), regard peer and teacher marks as being equivalent if the peer mark lies within one standard deviation of the tutor mark. Some studies do not make their definition of "agreement" explicit, and, consequently, such data were not analyzed. Future investigators would do well to avoid the use of proportions as a common metric.

Correlating common metrics with context variables

Thus, while displaying caution and awareness of over-enthusiastic interpretations of results by authors, some predictions deriving from published quantita-

TABLE 1

Summary of characteristics and common metrics of quantitative studies included in the meta-analysis

Study Identifiers	Population characteristics	Subject area and course name	What is assessed and level	Instrument & criteria[a]	Design quality[b]	Statistics reported	Value of common metrics[c]	Number involved in assessment
Billington (1997)	$n_{part} = n_{comp} = 58$ Final year undergraduates	Biology and Environmental Science Ecosystem Ecology	Advanced level Poster presentation skills	Evaluation sheet 5 grades over 5 criteria (TC) (G+)	H	Rank correlations Means	r = 0.80	3 faculty 57 students
Boud & Tyree (1979)	$n_{part} = n_{comp} = 28$ Female and male First year undergraduates	Law The Legal System	Introductory level Class participation	3 criteria Method (A) 10-point scale; Method (B) self-normalising scale (SC) (G+)	H	Product-moment correlation coefficients	mean r = 0.79	1 faculty 27 students
Burke (1969)	$n_{part} = n_{comp} = 12$ $n_{part} = n_{comp} = 20$ Graduates	Management Human Relations - application of behavioral science concepts/ theories	Introductory level Class Participation	Global rating, with some factors for consideration (G+)	L	Percentage agreement	83.3% 75.0%	1 faculty n(1) 11 students n(2) 19 students
Burnett & Cavaye (1980)	$n_{part} = n_{comp} = 186$ Fifth year undergraduate medical students	Medicine Surgery Clinical training & skills	Advanced level Peer performance	Global judgements considering 4 factors (TC) (G+)	H	Correlations	Mean r = 0.99	1 faculty 5 students
Butcher et al. (1995)	$n(1)_{part} = 28$; $n_{comp} = 7$ $n(2)_{part} = 27$; $n_{comp} = 9$ $n(3)_{part} = 32$; $n_{comp} = 8$ No gender information other than groups mixed 1st year undergraduates	Biology and Biochemistry Biosciences	Introductory level 3 group posters	Marking schedule with 3 criteria x 6 point scale (TC) (D)	H	Correlations Means and standard deviations	r(1) = 0.74 r(2) = 0.66 r(3) = 0.18 d(1) = 1.00 d(2) = 7.34 d(3) = -4.48	1 faculty Student numbers: n(1) = 4 n(2) = 3 n(3) = 4

TABLE 1 (cont.)

Study Identifiers	Population charactersitics	Subject area and course name	What is assessed and level	Instrument & criteria[a]	Design quality[b]	Statistics reported	Value of common metrics[c]	Number involved in assessment
Catterall (1995)	$n_{part} = n_{comp} = 105$ Part-time undergraduates	Business Marketing module	Introductory level Class test	Marking scheme designed by lecturers (TC) (D)	H	Means and standard deviations % agreement	d = 0.41 91% (<5% difference)	1 faculty 1 student
Chatterji & Mukerjee (1983)	(A) Graduate trainees $n_{part} = n_{comp} = 230$ (B) Graduate applicants $n_{part} = n_{comp} = 187$	(A) Engineering Group exercises (B) Hotel Management Group exercises	Introductory level Traits displayed in a group task and discussion.	12 criteria, 5-point scale Global assessments (G+)	H	Rank correlation	r(A) = 0.74 r(B) = 0.85	3 faculty c.7-12 students
D'Augelli (1973)	$n_{part} = n_{comp} = 168$ f = 85; m = 83 Undergraduates	Psychology Course not specified	Introductory level Interpersonal traits (GAIT) in dyadic interactions	Behavioral rating form 4 criteria (understanding, honest. warm, personally meaningful) 6 point scale (TC) (D)	H	Product moment correlations	Mean r = 0.29	2 faculty 5-7 students
Denehy & Fuller (1974)	$n_{part} = n_{comp} = 69$ (30 as evaluators) Freshmen	Dentistry Dental anatomy course	Introductory level Practical tests	List of criteria Point values on a normalised scale (TC) (G+)	H	Rank order correlations	r = 0.73 r = 0.71	2 faculty 3 students per set of projects
Eisenberg (1965)	$n_{part} = n_{comp} = 22$ Postgraduates	Psychology	Advanced level Examinations	Checklist rating scale Global rating by peers (C) (G+)	L	Spearman rank correlation	r = 0.84	1 faculty 2-9 students (asynchronous - prediction vs. end results)

TABLE 1 (cont.)

Study Identifiers	Population charactersitics	Subject area and course name	What is assessed and level	Instrument & criteria[a]	Design quality[b]	Statistics reported	Value of common metrics[c]	Number involved in assessment
Ewers & Searby (1997)	$n_{part} = n_{comp} = 71$ 2nd year undergraduates	Music Composition	Intermediate level Performance	Student defined criteria (SC) (G+)	H	Percentage agreement (derived)	52% (identical) 97% (<10% difference)	1 faculty ~ 5 students
Falchikov (1986)	$n_{part} = n_{comp} = 48$ 1st year undergraduates	Social science Cognitive psychology	Introductory level Essay	Marking schedule (AC) (D)	H	% agreement	60.6%	1 faculty 1 student
Falchikov (1994)	$n_{part} = 101$; n_{comp} = approx 24 Groups mixed in terms of gender First year undergraduates	Social Science Foundations of social science	Introductory level Oral presentations	Peer assessment form (AC) (G+)	H	Means and standard deviations	d = 0.00	1 faculty 15 students (av)
Falchikov (1995)	$n_{part} = n_{comp} = 13$ (f = 12; m = 1) Third year undergraduates	Biological Science Developmental Psychology	Introductory level Oral presentations	Student generated criteria (SC) (D)	H	Percentage agreement	97.5% Agr < + 0.5 out of 20 marks	1 faculty 12 students
Falchikov (1999)	$n_{part} = n_{comp} = 42$ (f = 39; m = 3) Second year undergraduates	Social science Lifespan development	Introductory level Self-assessment test	Marking schedule (TC) (D)	H	Means and standard deviations Correlation	d = 0.23 r = 0.92	1 faculty 1 student
Fineman (1981)	$n_{part} = n_{comp} = 12$ (f = 4; m = 8) Third year undergraduates	Business Administration Organisational Behavior	Advanced level Various: experiential exercises; group discussion; class experiments	5 criteria (SC+TC) (D)	H	Means and standard deviations for 5 criteria	Mean d = 0.17	1 faculty 11 students

TABLE 1 (cont.)

Study Identifiers	Population charactersitics	Subject area and course name	What is assessed and level	Instrument & criteria[a]	Design quality[b]	Statistics reported	Value of common metrics[c]	Number involved in assessment
Freeman (1995)	n_{part} = 210, but qualitative data for 39 3 sessions $n(1)_{part}$ = 17; n_{comp} = 3 $n(2)_{part}$ = 11; n_{comp} = 2 $n(3)_{part}$ = 11; n_{comp} = 2 Final year undergraduates	Business degree Securities marketing regulation	Advanced level Oral Presentations	22 point checklist (TC) (C) (D)	H	Means and standard deviations for the three sessions and total Correlation	d (1) = -0.70 d (2) = 1.28 d (3) = -0.31 r = 0.6	2 faculty 4-6 students
Friesen & Dunning (1973)	n_{part} = n_{comp} = 12 (f = 5; m = 7) Master's level graduates	Practicum in a guidance counselling program	Advanced level Analysis of videotaped interviews	'Rating scale of counselor effectiveness' checklist (TC) (D)	H	Rank order correlation (rho)	r = 0.88	5 faculty 12 students
Fry (1990)	n = 70, but quantitative data reported for only n_{part} = n_{comp} = 11 First year undergraduates	Mechanical Engineering Mechanics	Introductory level Tutorial problems	Marking scheme with model solutions (TS) (D)	H	Raw data converted to means and standard deviations & correlation	d = -0.29 r = 0.87	1 faculty 1 student?
Fuqua, Johnson, Newman, Anderson, & Gade (1984)	n_{part} = n_{comp} = 36 (f = 21; m = 15) Postgraduates	Pre Practicum counselling methods course	Mixed level Counselling practice with coached clients	Standardised rating format 6 dimensions on a 5-point scale (TC) (D)	H	Pearson product-moment correlation	Mean r = 0.66	3 supervisors "Small groups" of students
Gray	n_{part} = n_{comp} = 96 Undergraduates	Engineering materials course	Introductory level Exam paper marking	Model solution (TS) (D)	H	Frequency graph with % agreements	39% (identical ratings) 58% (+/-10% estimated)	1 faculty 1 student

TABLE 1 (cont.)

Study Identifiers	Population charactersitics	Subject area and course name	What is assessed and level	Instrument & criteria[a]	Design quality[b]	Statistics reported	Value of common metrics[c]	Number involved in assessment
Hammond & Kern (1959)	$n(A)_{part} = n_{comp} = 77$ $n(B)_{part} = n_{comp} = 70$ $n(C)_{part} = n_{comp} = 60$ 3 cohorts of undergraduates yrs 1-4, (A)1954, (B)1955,(C)1956	Medicine Competence as a physician	Introductory to intermediate to advanced levels	Global ratings (G) Part of a much larger study investigating relationships between individual attributes, peer judgements and performance	L	Correlation coefficients	Mean r(A)= 0.35 Mean r(B) = 0.45 Mean r(C) = 0.54	? faculty student numbers n(A) = 76 n(B) = 69 n(C) = 59
Hughes & Large (1993)	$n_{part} = n_{comp} = 44$ Final year undergraduates	Pharmacology Communication skills	Advanced level Oral presentation skills	Mark out of 100% N of discri. = 50 (Faculty only) (AC) (G) (students)	H	Correlation coefficient Means and standard deviations	r = 0.83 d = -0.43	7 faculty 43 students
Hunter & Russ (1996)	$n_{part} = n_{comp} = 23$ (A) Year 2 undergraduates (B) Final year undergraduates	Music Performance studies	(A) Introductory level (B) Advanced level Performance	Criteria agreed and checklist provided Criteria for classification provided (AC) (TC) (G+)	H	Raw scores transformed into correlation coefficients & Means & standard deviations	Mean r = 0.77 Mean d = 0.15	"A panel" of faculty (*n* not specified) Student panels n = 6 or 7
Jacobs, Briggs & Whitney (1975)	$n_{part} = n_{comp} = 67$ Sophomores	Dentistry Orthodontic course	Intermediate level Assessment of examination paper	Global score, 4-point scale of given criteria (TC) (G+)	H	Pearson product moment correlation coefficient for instructors vs. stratified groups of peers	r (high GPA) = 0.58 r (above av.) = 0.48 r (below av.) = 0.56 r (low GPA) = 0.27	2 faculty 67 students (each given 4 papers, one from each quartile of GPA score)

TABLE 1 (cont.)

Study Identifiers	Population charactersitics	Subject area and course name	What is assessed and level	Instrument & criteria[a]	Design quality[b]	Statistics reported	Value of common metrics[c]	Number involved in assessment
Kaimann (1974)	$n_{part} = n_{comp} = 25$ Postgraduates	Business Administration Production Management	Advanced level? Oral presentation skills and critiquing skills	Global discrimination (G)	L	Spearman's rank correlation	r = 0.84	1 faculty 24 students
Kegel-Flom (1975)	$n_{part} = m = n_{comp} = 110$ Medical interns	Medicine	Introductory level Intern performance	Rating on 4 dimensions over 12 scores (D)	L	Correlations - peer rating and years 1, 2, 3, and 4 grades	Mean r = 0.25	Not stated
Kelmar (1992)	$n(1)_{part} = n_{comp} = 12$ $n(2)_{part} = n_{comp} = 20$ $n(3)_{part} = n_{comp} = 27$ Graduates	Management General management	Level? Oral presentation skills	Global rating (AC) (G+)	H	Means and standard deviations Correlations (derived from raw data)	d(1) = 0.51 d(2) = 0.62 d(3) = 0.41 r(1) = 0.53 r(2) = 0.78 r(3) = 0.77	1 faculty 11/ 19/ 26 students
Korman & Stubblefield (1971)	$n_{part} = n_{comp} = 68$ Senior year medical students	Medicine	Advanced level Comparing past experience and grades to internship performance	Peer ratings on 12 variables compared with 8 internship characteristics (TC) (D)	L	Correlation coefficient	r = 0.14	5 + 7 faculty 67 students

TABLE 1 (cont.)

Study Identifiers	Population charactersitics	Subject area and course name	What is assessed and level	Instrument & criteria[a]	Design quality[b]	Statistics reported	Value of common metrics[c]	Number involved in assessment
Kwan & Leung (1996)	$n_{part} = n_{comp} = 96$ 3rd year Higher Diploma students	Hotel Personnel and Training course	Advanced level Simulation training exercise	Checklist 12 dimensions on a 6 point scale (TC) (C) (D)	H	Means and standard deviations Percentage agreement and Correlation coefficient	d = -0.11 69.7% r = 0.48	1 faculty 15-18 students 96 pairs of results compared
Lennon (1995)	$n_{part} = n_{comp} = 49$ 2nd year undergraduates (1) = peer model (2) = peer observer	Health Sciences Physiotherapy	Introductory level Practical simulation	Marking scheme Criteria (AC) (D)	H	Correlations % agreement	r(1) = 0.34 r(2) = 0.55 (1) 87% (2) 83% (<10 marks difference)	1 faculty 1 student
Linn, Arostegui & Zeppa (1975)	$n_{part} = n_{comp} = 54$ Junior medical students	Medicine	Advanced level Performance on ward assignments, compared to final grades	Performance rating scale 16 items on a 4 point scale (TC) (D)	H	Correlation coefficient	mean r = 0 .42	? faculty 7-10 students
Magin (1993)	$n_{part} = n_{comp} = 169$ 1st year undergraduates	Medicine (1) Introductory clinical and behavioral studies (2) Human behavior	Introductory level Group process, final report and presentation	2 criteria on 5 point scale (AC) (G+)	H	Product moment correlation	r(1) = 0.85 r(2) = 0.79	1 faculty c. 8-10 students
Magin & Churches (1988)	$n_{part} = n_{comp} = 87$ Second level Undergraduates	Mechanical Engineering Practical Design	Intermediate level Examination paper	Marking schedule (TC) (TS) (D)	H	Means and standard deviations Product moment correlation	d = -0.37	1 faculty 1 student

TABLE 1 (cont.)

Study Identifiers	Population charactersitics	Subject area and course name	What is assessed and level	Instrument & criteria[a]	Design quality[b]	Statistics reported	Value of common metrics[c]	Number involved in assessment
Melvin & Lord (1995)	n_{part} = 410 n(Psychology) = n_{comp} = 57 n(Accounting) = n_{comp} = 83 n(Finance) = n_{comp} = 51 n(Marketing) = n_{comp} = 74 n(Gen business) = n_{comp} = 145 7 graduate courses 11 undergraduate courses	Psychology Accounting Finance Marketing General business	Mixed levels Quality & quantity of class participation	Global rating, 3 point scale with faculty subdivisions and with forced distribution (G)	H	Pearson's or point biserial correlation	r(P) = 0.94 r(A) = 0.89 r(F) = 0.71 r(M) = 0.75 r(GB) = 0.67	11 faculty student numbers n(P) = 7 - 17 n(A) = 8 - 22 n(F) = 4 - 32 n(M) = 13 - 38 n(GB) = 11
Mont-gomery (1986)	n_{part} = n_{comp} = 54 Groups "balanced for gender" Undergraduates	Basic Communica-tion	Introductory level Exhibited openness in group discussion	Openness rating criterion 7-point scale (TC) (G)	H	Correlations for 2 discussion groups	r = 0.71 r = 0.79	3 faculty (trained observers) 5 students
Morton & Macbeth (1977)	n_{part} = n_{comp} = 138 First clinical year medical undergraduates	Medicine Surgery	Introductory level Clinical perfor-mance	Assessment form Global judgement (peers) (faculty assessment grade includes exam papers and a presentation) (G)	L	Means and standard deviations Correlation	d = -0.16 r = 0.53	4 faculty 4-5 students
Mowl & Pain (1995)	n_{part} = n_{comp} = 53 First Year Undergraduates	Geography Human Geo-graphy	Introductory level Essay	Peer assessment form Global assessment (AC) (G+)	H	Means and standard deviations Correlation coefficient	d = -0.46 r = 0.22	1 faculty 1 student
Ngu et al. (1995)	n_{part} = n_{comp} = 17 Master's level students	Computing Science and Engineering Advanced database management	Advanced level Essay questions	The *Peers* computerised system (AC) (G+)?	H	Means and standard deviations	d = 0.18	1 faculty 2 students

TABLE 1 (cont.)

Study Identifiers	Population charactersitics	Subject area and course name	What is assessed and level	Instrument & criteria[a]	Design quality[b]	Statistics reported	Value of common metrics[c]	Number involved in assessment
Oldfield & Macalpine (1995)	n (1) $_{part}$ = n $_{comp}$ = 18 1st year undergraduates n (2) $_{part}$ = n $_{comp}$ = 12 2nd year undergraduates n (3) $_{part}$ = n $_{comp}$ = 47 2nd yr. Higher Diploma students	Engineering	Introductory level Intermediate level Intermediate level Contribution of peers: brief lecture, report and/or essay, group work	5 or 9 point scale (G)	L	Correlation coefficients	r(1)= 0.16 r(2) = 0.72 r(3) = 0.91	1 faculty student group - 1 (size not known)
Orpen (1982)	n (A) $_{part}$ = n $_{comp}$ = 21 2nd year undergraduates n (B) $_{part}$ = n $_{comp}$ = 21 3rd year undergraduates	Organizational behavior Political philosophy	Intermediate level Essay Advanced Essay	3 criteria Global judgement (6 point scale) (TC) (G+)	L	Means and mean variances	d(A) = 0.20 d(B) = -0.03	5 faculty 5 students
Orsmond, Merry & Reiling (1996)	n $_{part}$ = 78; n $_{comp}$ = 39 Undergraduates	Science Comparative Animal Physiology	Introductory level Group poster presentation	Marking form 5 criteria on a 5 point scale (TC) (D)	H	Percentage agreement Spearman's rank	18% (identical marks) r = 0.73	1 faculty c. 39 students
Pease (1975)	n $_{part}$ = n $_{comp}$ = 60 (f = 55; m = 5) Undergraduates	Sociology, History, English Teacher education	Level varied: data collected over 2 year period Teacher performance	Single numerical rating on a 10 point scale (G)	H	Rank correlation	r = 0.82	1 faculty 30-33 students

TABLE 1 (cont.)

Study Identifiers	Population charactersitics	Subject area and course name	What is assessed and level	Instrument & criteria[a]	Design quality[b]	Statistics reported	Value of common metrics[c]	Number involved in assessment
Ritter (1997)	$n_{part} = n_{comp} = 136$ Trainee teachers	Trainee primary teaching course Introductory Australian history	Introductory level Group participation	Two criteria, 4 grade scale (SC) (G+)	H	Percentage agreement	67%	1 faculty 15-20 students
Rushton, Ramsey & Rada (1993)	$n_{part} = n_{comp} = 32$ Final year undergraduates	Computer Science	Advanced level Collaborative authoring of essay	Five criteria on a 10-point scale (TC) (D)	H	Percentage agreement	62.5% Agr = differ by 9% or less	1 faculty 2-3 students
Stefani (1992)	$n_{part} = n_{comp} = 63$ 1st year undergraduates	Biochemistry Laboratory practical experiment	Introductory level Laboratory report	Marking schedule with 6 categories (SC) (D)	H	Percentage agreement	14.03% identical 80.70% +/- 10 marks	1 faculty 1 student
Stefani (1994)	$n_{part} = n_{comp} = 57$ Undergraduates	Biological Sciences Biomedical techniques	Introductory level Laboratory report	Student defined marking schedule (SC) (D)	H	Means and standard deviations Correlation coefficient	d = 0.04 r = 0.89	1 faculty 1 student
Wiggins & Blackburn (1969)	$n(1)_{part} = n_{comp} = 46$ 1st year graduates $n(2)_{part} = n_{comp} = 58$ 1st year graduates	Psychology	Introductory level Academic prediction (of cumulative GPA) Introductory level Academic prediction	Checklist 16 variables (performance and personality traits) (TC) (D)	H	Correlations between 16 variables and cum GPA averaged	Mean r(1) = 0.25 Mean r(2) = 0.35	Cumulative GPA score 45 students Cumulative GPA score 57 students

[a] AC = agreed criteria; D = dimensional judgement; G = global judgement; G+ = global considering some aspects; SC = student criteria; TC = tutor criteria (including checklists); TS = tutor's solution

[b] H = high quality study; L = low quality study

[c] % refers to percentage agreement; d = effect size; r = correlation coefficient.

Note: n_{part} = number of participants; n_{comp}= number of comparisons.

tive peer assessment studies may be made. The following variables have been identified by researchers as mediating the correspondence between faculty and peer ratings: ability of student raters (Jacobs at al, 1975); practice effects (Orpen, 1982; Fuqua, Johnson, Newman, Anderson, & Gade, 1984; Hunter & Russ, 1996); number of student raters related to the reliability of marking (Magin, 1993); methodologies employed (Falchikov, 1986); the type of assessment involved (Mowl & Pain, 1995). In addition, Falchikov and Boud's (1989) meta-analysis identified the following significant variables in the context of student self assessment: the level of course; the complexity of measurements used; the explicitness of criteria and student ownership of these; the subject area in which the assessment takes place. Given the similarities between self and peer assessment, it is likely that these variables may influence peer assessment outcomes, too.

In the light of previous research, it may be hypothesized that

1. There will be subject area differences in the validity of peer assessment (defined as the similarity between peer and faculty marks), with higher validities being associated with science and engineering areas than with social science and arts (Falchikov & Boud, 1989).
2. Peer assessment carried out in advanced level courses will be more valid than that in introductory courses (Falchikov & Boud, 1989).
3. The greater the number of students involved in each peer assessment, the better the correspondence between peer and teacher marks (c.f. Magin, 1993).
4. Explicit and student owned criteria will be associated with better peer assessment validities than other criteria or absence of criteria (Fineman, 1981; Falchikov, 1986; Stefani, 1994).
5. The nature of the assessment task will influence validity of peer assessment, with assessments carried out in traditional academic areas within the classroom (e.g., essays, tests, presentations) having better validities than those in areas of professional practice (e.g., intern performance, counselling skills, teaching practice).
6. More valid assessments will be associated with higher quality studies than those deriving from studies with poor experimental designs (Falchikov & Boud, 1989).
7. Different levels of teacher-peer correspondence will result from studies where ratings involve making decisions in different ways (e.g., where large or small numbers of dimensions are involved or where familiar ranges (e.g., 100%) are used). Some researchers argue that ratings based on large numbers of dimensions will lead to closer correspondence between peer and teacher ratings (Harris & Schaubroeck, 1988), while others argue the reverse case (Falchikov & Boud, 1989).

Results

Correlation coefficients were reported in 56 experimental conditions. Some studies reported more than one *r* value. In cases where these were dependent, an average was calculated (there being insufficient information for the calculation of a weighted average). In other cases, the r values were independent (e.g., Chatterji & Mukerjee, 1983). Distribution of independent r values is shown in Figure 1.

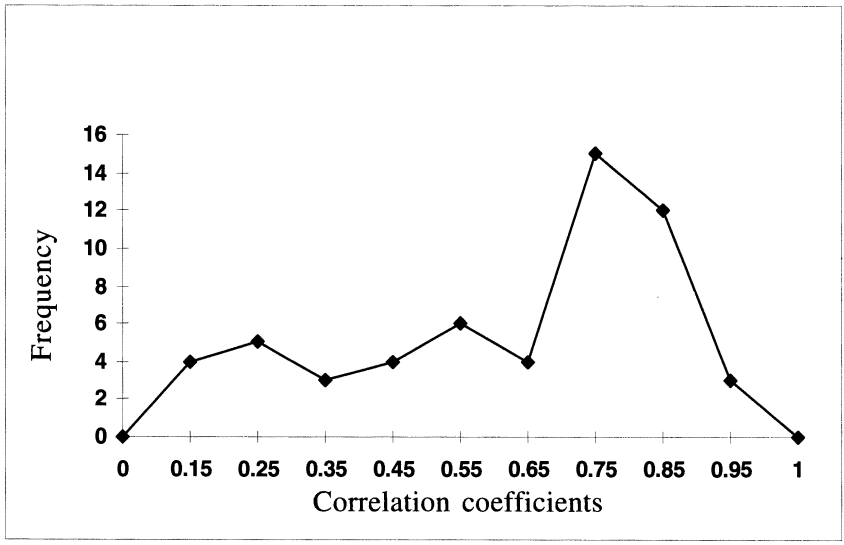

FIGURE 1. *Distribution of correlation coefficients*
Note: Points on the x axis represent the mid point of a range of correlation coefficient values. For example, 0.25 is the mid point of the .20 to .29 range

In the present meta-analysis, values of r varied from 0.14 to 0.99. The mean overall value was $r = 0.69$, calculated by converting the r values to z scores and weighting them by the usual weights, n_i-3, where n_i is the number of comparisons in study i. (Correlations are very dependent on study size, so these weights allow correlations based on larger studies to be given greater weight. For a discussion of the rationale behind the conversion and the weights, see Shadish and Haddock, 1994, pp. 265-269.) This is a very significant average value of r, suggesting that overall peer marks agree well with teacher marks.

A test for homogeneity in the correlation coefficients was carried out using z scores, and significant heterogeneity was evident ($Q = 1036$, $df = 55$, $p < .001$). One study (Burnett and Cavaye, 1980) showed an usually large z score of 2.65 despite being a large scale study ($n = 186$). With the Burnett and Cavaye study excluded, the Q value reduced to 469, still indicating significant heterogeneity ($p < 0.001$). Apart from this study, the distribution of z values showed no other irregularities that would indicate any publication bias.

A Cautionary Note About the Burnett and Cavaye (1980) Study

The study by Burnett and Cavaye (1980) reported an almost perfect correlation between peer assessment and final grade ($r = 0.99$). This is particularly surprising given the large number of students involved ($n = 186$) and the fact that the authors claimed that instructions for "explicit or constructive use of the criteria" (p. 274) were avoided. However, in this study peer assessment percentage *marks* were correlated with the overall *grade* (from a seven point scale made up of bands of percentage marks, e.g., 75% to 84%), a practice which could act to increase the degree of agreement. Moreover, peer assessment, which was

related to performance within a small group over a period of weeks prior to the examination, included attendance as a criterion. This is an unambiguous criterion, as absence or presence is easily observed by all and often recorded, and is, thus, also likely to increase agreement between raters.

Effect size calculations derived from 24 experimental conditions show a weighted mean value of $d = 0.24$ (see Figure 2). The weights used are the usual weights for effect sizes, in other words, $1/u_i$, where $u_i=(n^f_i+n^p_i)/n^f_i n^p_i + d_i^2/2(n^f_i+n^p_i)$, n^f_i and n^p_i being the number of faculty and the number of peers assessing each student in study i, respectively, and d_i the effect size in study i (Shadish & Haddock, 1994, p 268). Dependent effect sizes within one study were averaged as above.

The range of d values, from $d = -4.48$ to $d = 7.34$ indicates both some peer over marking (positive values of d) as well as some under marking (negative values) compared with teachers. However, the two extreme values derive from the same study (Butcher, Stefani, & Tariq, 1995). There is some reason to regard this study as atypical—the range of d values otherwise is from -0.75 to 1.25 with a weighted mean of -0.02.

A Cautionary Note About the Butcher, Stefani, & Tariq (1995) Study

The possibility that the study by Butcher et al (1995) may be atypical was investigated. This study compared faculty and peer marks awarded for poster presentations of topics in the biosciences and produced poor faculty-student correspondences. There are a number of factors which, together, might give rise to this result.

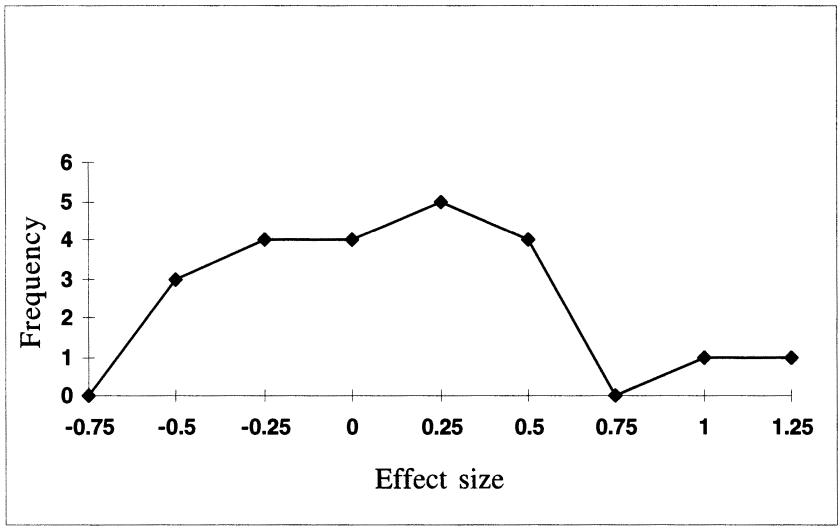

FIGURE 2. *Distibution of effect sizes*

Note: Two extreme effect size values were also found: 7.34 and -4.48. Points on the x axis represent the mid point of a range of effect sizes. For example, 0.4 is the mid point of the 0.2 to 0.6 range.

1. The reporting of this study gives rise to confusion regarding the conduct and location of the peer assessments.
2. While student peer means were similar for the three topics, those for faculty differed considerably, varying between 56.1 and 77.7. In addition, peer standard deviations exceeded or equalled those for faculty in all cases.
3. The arrangements for assessment of posters are unclear. Students were reported to have assessed all posters. Thus, assuming roughly equal numbers of posters at each geographical location, each student would have assessed nine posters in one afternoon. Effects of boredom or tiredness cannot be discounted.
4. Poster projects were cross curricular, "each requiring knowledge and understanding of several branches of the biosciences" (p. 166). This feature, while admirable in itself, is likely to increase student assessment difficulties, and reduce peer-teacher agreement, in that beginner students are being asked to come to terms with more than one new discipline.
5. Another feature which is thought to be associated with poorer teacher-peer agreement was also present. Teachers supplied the assessment criteria, but it is not clear whether all teachers involved at all locations were party to the identification of criteria. It is possible that the decision relating to criteria was taken by one collaborator alone. Thus, lack of understanding of the implicit elements contained within the list of criteria could extend to teachers at two of the three sites as well as to students.

This study may be categorized as an outlier, in that effect size homogeneity statistics indicated a change when data from the Butcher et al study were excluded (cf., Hedges & Olkin, 1985). The Q value for all effect sizes was 92.29. This reduced to 2.56 when data from the Butcher study was excluded from the calculation. Omitting the Butcher data, there was no reason to doubt homogeneity of effect sizes ($p = 0.99$). Thus, results will be calculated both including and excluding data from the Butcher et al. (1995) study.

The weighted mean effect size when the Butcher study is omitted is -0.02, indicating no consistent disagreement between faculty and peers on average.

Methods of Analysis

Weighted multiple regression models were built both for the correlations and for the effect sizes in order to investigate the dependence on the context variables. For the correlation, the Fisher transformation $z = 0.5\ln [(1+r)/(1-r)]$ was used as the response variable (Shadish & Haddock, 1994, p. 268). The weights used were those previously discussed; they place more weight on the larger studies.

Qualitative variables such as "Subject area" were converted to indicator variables. This was also done for variables such as "Level of course," which presented an ordinal level of measurement.

First, the effect of each context variable was examined individually by means of a regression model involving only that variable as a predictor. The p-value for the hypothesis of 0 slope was calculated to test for a significant difference

between the variable levels. Combinations of context variables were then tried as predictors in best-subsets regression to determine what combinations had significant influences. Finally, a multivariate model that explained as much as possible of the variation was sought using stepwise procedures. The statistical package MINITAB was used to calculate the models and their goodness of fit by

a) determination of the adjusted R-squared value
b) determination of the t and p-values for each predictor variable in the models
c) calculation of variance inflation factors for each predictor variable to check for multicollinearity between the variables
d) performance of data subsetting lack-of-fit tests to test for curvature and interactions between the variables

Correlations

The unusual study by Burnett and Cavaye discussed previously was flagged by the analysis as having a large influence on the models of the correlation. The analyses were repeated without this study to see if the significant factors remained significant for the remaining studies.

For the full set of studies, including that of Burnett and Cavaye (1980), the factors which had a significant effect on the correlation coefficient individually were as follows:

- Dimensionality vs. Global judgements ($p < 0.001$)
- Nature of assessment task ($p < 0.001$)
- Quality of study ($p < 0.01$)
- Number of peers ($p < 0.02$)
- Level of course ($p < 0.04$)

With the Burnett and Cavaye study omitted, the level of the course is no longer significant but the subject area becomes significant ($p<0.025$). The study in question was at an advanced level in medicine. Omitting it removes the significant effect of advanced level courses, but reduces the average r for medical subjects to significantly below other subject areas. Involvement of students in deciding the criteria for assessment becomes marginally significant ($p<.07$). Table 2 summarises the significance of the context variables, on their own, in explaining variation in the common metrics. Significant values are highlighted. However, it must be remembered that each of these analyses is looking at one factor in isolation. Some of the observed effect may be due to the influence of other variables, even though the multicollinearity between most of them was small. To overcome such potential interaction problems, a multiple regression analysis was performed which looks at the effect of each factor while controlling for other factors.

Effect Sizes

When the Butcher study is included, none of the factors had a significant influence on the effect size, d_i. Even when Butcher is omitted, only the number

TABLE 2
Significance of context variables

Variable	*p* value for *r*	*p* value for *r* without Burnett & Cavaye	*p* value for d	*p* value for *d* without Butcher
Quality of study	**0.01**	**0.005**	0.89	0.84
Subject area	0.38	**0.025**	0.63	0.11
Nature of task	**0.001**	**0.001**	0.97	0.92
Level	**0.04**	0.49	0.79	0.88
Number of peers	**0.02**	**0.02**	0.33	**0.09**
Dimens versus Global	**0.000**	**0.000**	0.48	**0.06**
Criteria (SC/TC)	0.83	**0.07**	0.65	0.66

of peers and whether the judgements were global or by dimension show any statistical significance. No significant combinations of factors were found.

The different effect of the context variables on the two common metrics is not surprising, nor is the fact that the correlation coefficients display heterogeneity while, when Butcher is excluded, the effect sizes display homogeneity. The smaller number of studies reporting effect sizes makes it harder to draw any inferences than when analyzing the correlation coefficients. Effect size is also measuring something very different from correlation. A "perfect" effect size in this context ($d_i = 0$) would require peers and faculty to give a piece of work the same mark on average. A perfect correlation ($r = 1$) requires peers and faculty to be able to agree on where the piece of work sits on a scale but can be achieved with the peer average mark very different from the faculty average mark. A context variable that improves how good students are at ranking their peers may not reduce their d value at all. In addition, effect size is averaged over all the pieces of work looked at in a study, and a piece of work marked much higher by peers than faculty can be balanced by another piece of work marked much higher by faculty than by peers. Perfect correlation requires the relationship between peer mark and faculty mark to be the same for all pieces of work. Thus it is possible that a context variable that is related to less reliability in marks between pieces of work may have more influence on correlation coefficients than on effect size.

We now proceed to interpret these results in the light of our hypotheses, starting with a comparison of the type of judgements required. We then look at the nature of the assessment task and investigate its effect on outcomes, before moving on to investigate the effects of study design quality, the number of peers involved in each assessment, subject area differences, the status of criteria and level of course comparisons.

Dimensionality Versus. Global Judgements

Statistics relating to studies where peers were required to make an overall global judgement (G) with no explicit criteria were compared with those where overall judgements entailed consideration of several dimensions or criteria (G+).

TABLE 3
Dimensionality versus Global judgements: mean values

(weighted means)	G	G+	D	*p* values
Mean *r*	0.72 (n = 17)	0.85 (n = 19)	0.53 (n = 18)	*0.000*
Mean *r* omitting Burnett & Cavaye	0.72	0.77	0.53	*0.000*
Mean *d*	-0.32 (n = 2)	0.17 (n = 10)	0.34 (n = 13)	*0.48*
Mean *d* omitting Butcher	-0.32	0.17	0.03	*0.06*

In addition, in some studies, students were asked to make judgements for each dimension separately (D). These data are shown in Table 3.

The D category produces a significantly lower mean correlation between faculty and students than the G and G+ categories, with the G+ correlations also significantly higher than those from G ratings when the Burnett & Cavaye study is included. G ratings seem to be associated with a larger effect size representing poorer peer-teacher agreement than the other types (although based on only two G studies). Thus, the intermediate G+ category, where judgements are made in the knowledge of criteria or guidelines, may give rise to slightly better peer-faculty agreement than the two other categories.

Nature of Assessment Task Comparisons

Statistics deriving from studies which involved peer assessment of professional practice (e.g., clinical skills, teacher performance) were compared with those deriving from assessment of traditional academic activities: academic products (e.g., essays, examinations) and academic processes (e.g., oral presentation

TABLE 4
Nature of assessment task comparisons: mean values

(weighted means)	Professional Practice	Academic Product	Academic Process	*p* values
Mean *r*	0.54 (n = 15)	0.75 (n = 14)	0.83 (n = 25)	*0.001*
Mean *r* omitting Burnett & Cavaye	0.53	0.75	0.76	*0.001*
Mean *d*	-0.02 (n = 3)	0.32 (n = 11)	-0.05 (n = 10)	*0.97*
Mean *d* omitting Butcher	-0.02	0.03	-0.05	*0.92*

skills, participation in group activities). A summary of relevant statistics may be found in Table 4.

Mean correlation coefficients indicate that peer assessment in the area of professional practice ($r = 0.54$) may be more problematic than in either of the academic areas, where correlations are $r = 0.75$ and $r = 0.83$ (0.75 and 0.76 with Burnett & Cavaye excluded). Mean effect sizes do not differ significantly between the three groups.

Study Design Quality

Studies rated as having high design quality were contrasted with those of low quality. Summary data are shown in Table 5.

Only nine of the studies were rated as being low quality. These contributed 11 independent comparisons. Seven of the nine low quality studies were conducted and published over twenty years ago and only one low quality study was published in the 1990s. The mean r value for low quality studies ($n = 11$) is 0.5, significantly lower than the mean value of 0.78 for high quality studies ($n = 45$), or 0.72 if Burnett & Cavaye is excluded.

The mean effect size (excluding Butcher) is not significantly different between high and low quality studies, but only three of the latter reported effect sizes.

Number of Peers Involved in Each Assessment

Peer assessments were noted to have been carried out by a varying number of students throughout the corpus and by relatively few singletons. Mean statistics were compared for four group sizes (1 student, 2-7 students, 8-19 students and 20+ students per assessment) although regression analysis was also performed for the ungrouped data where those were available. Results are shown in Table 6.

The correlations were significantly smaller as the number of peers increased, based on the ungrouped data. Group sizes of 20 or more produced mean corre-

TABLE 5
Design quality: mean values

(weighted means)	High Quality	Low Quality	p values
Mean *r*	0.78 ($n = 45$)	0.50 ($n = 11$)	*0.01*
Mean *r* omitting Burnett & Cavaye	0.72	0.50	*0.005*
Mean *d*	0.25 ($n = 21$)	0.01 ($n = 3$)	*0.89*
Mean *d* omitting Butcher	-0.03 ($n = 18$)	0.01	*0.84*

TABLE 6
Number of peers involved in each assessment: mean values

(weighted means)	1	2 - 7	8 - 19	20+	*p* values
Mean *r*	0.72	0.81	0.77	0.59	*0. 02*
	(*n* = 7)	(*n* = 12)	(*n* = 12)	(*n* = 15)	
Mean *r* omitting Burnett & Cavaye	0.72	0.59	0.77	0.59	*0.02*
Mean *d*	-0.07	0.43	0.24	-0.31	*0.33*
	(*n* = 6)	(*n* = 11)	(*n* = 5)	(*n* = 2)	
Mean *d* omitting Butcher	-0.07	0.05	0.24	-0.31	*0.09*

lation coefficients significantly lower than the rest of the studies as a whole. Ratings by singletons do not appear to be less reliable than others. The mean effect size for larger groups is also larger in absolute value than is the effect size observed with smaller groups.

Subject Area Differences

Studies comprising the quantitative peer assessment studies database came from a wide variety of subject areas (see table 1). Studies have been categorized into the following groups:

- Business and Management (including Hotel Management)
- Medicine, Dentistry, and paramedical subjects
- Science and Engineering
- Social Science and Arts

Summary data are shown in Table 7.

Correlation coefficients did not differ significantly between subject areas when the Burnett and Cavaye study was included, but, when this study was omitted, the studies in the area of Medicine and Dentistry recorded significantly lower

TABLE 7
Subject area differences: mean values

(weighted means)	Business Management	Medicine, Denistry, & Paramedical	Science & Engineering	Social Sciences & Arts	*p* values
Mean *r*	0.74	0.74	0.76	0.66	*0.38*
	(*n* = 11)	(*n* = 19)	(*n* = 12)	(*n* = 14)	
Mean *r* omitting Burnett & Cavaye	0.74	0.57	0.76	0.66	*0.005*
Mean *d*	0.20	-0.36	0.49	0.02	*0.63*
	(*n* = 10)	(*n* = 2)	(*n* = 7)	(*n* = 5)	
Mean *d* omitting Butcher	0.20	-0.36	-0.09	0.02	*0.11*
			(*n* = 4)		

correlations than the other areas. Correlations in Social Science and Arts were slightly lower, but not significantly so.

Effect sizes for Medicine and Dentistry showed peers significantly under-marking compared to faculty, and peers in Business and Management significantly over-marking. However, there were only two studies reporting effect sizes in Medicine and Dentistry and ten in Business and Management, and the variable "Subject" is not a significant predictor as a whole.

Status of Criteria

Table 8 shows mean values for the two types of criteria: criteria where students have been involved in their selection (SC, student derived criteria, and AC, agreed criteria) and criteria devised by tutors only (TC, tutor's criteria, and TS, tutor's solution).

When the Burnett and Cavaye study is omitted, student derived and agreed criteria are fairly significantly associated with better teacher-peer agreement than the teacher supplied criteria group. Fewer studies were included in this analysis as the criteria "status" is not reported for 18 studies.

Level of Course Comparisons

Assessments associated with "Introductory" (year 1), "Intermediate" (year 2), and "Advanced" (year 3 and above) courses were compared. Summary statistical data are shown in Table 9.

In terms of the mean correlation coefficients, peer assessment agrees progressively more with faculty markers as the course level increases when Burnett and Cavaye is included. This significant difference is not, however, evident in the remaining studies nor in the effect sizes. It appears to result primarily as a result of the very high correlation Burnett and Cavaye achieved.

Interactions Among Variables

Investigation of interactions between pairs (or more) of variables revealed no surprises, with no variable's behaviour changing significantly when combined with other variables. However, the number of studies in some of the subsets was

TABLE 8
Status of criteria: mean values

(weighted means)	SC & AC	TC & TS	*p* values
Mean *r*	0.75 ($n = 12$)	0.76 ($n = 26$)	*0.83*
Mean *r* omitting Burnett & Cavaye	0.75	0.59	*0.07*
Mean *d*	-0.05 ($n = 9$)	0.29 ($n = 14$)	*0.65*
Mean *d* omitting Butcher	-0.05	0.04	*0.66*

Note: SC = students' criteria; AC = agreed criteria; TC = teacher's criteria; TS = teacher's solution.

TABLE 9
Effects of level of course: mean values

(weighted means)	Introductory level	Intermediate level	Advanced level	*p* values
Mean *r*	0.67 (n = 28)	0.77 (n = 8)	0.87 (n = 13)	0.04
Mean *r* omitting Burnett & Cavaye	0.67	0.77	0.64	0.49
Mean *d*	0.36 (n = 10)	0.11 (n = 3)	-0.05 (n = 8)	0.79
Mean *d* omitting Butcher	-0.01	0.11	-0.05	0.88

often very small. The reduction in the correlation when the subject area was Social Science and Arts did, nevertheless, become significant in the model including Burnett and Cavaye once the other variables of advanced level courses, dimensional criteria, professional practice assessment, and large numbers of peers had been taken into account.

The "best" multivariate model for r for the full data set had an F-test p-value of $p < 0.001$, an adjusted R-squared value of 66%, and all predictors significant, yet with no evidence from the various tests to suggest lack of model fit. However, because of missing information in some studies (such as unknown Level), the model was based on only 39 cases. The regression equation for z is

$$z = 1.63 - 0.759 \text{ prop} - 0.656 \text{ ss} - 0.449 \text{ D} - 0.877 \text{ np4} + 0.677 \text{ lvl3},$$

where prop is 1 for a professional practice task and 0 otherwise; ss is 1 for a social science subject area and 0 otherwise; D indicates studies where students were required to make judgements for each dimension separately; np4 is 1 for more than 20 peers and 0 otherwise; and lvl3 is 1 for an advanced course and 0 otherwise.

When the Burnett and Cavaye study is omitted, the new model that best fits the remaining data no longer includes number of peers, advanced level courses or social science as significant factors. It has an adjusted R-squared of only 45%, but is based on 53 cases. Being a high quality study is now included as the only positive factor. (Quality can be included as an extra significant factor in the model which includes Burnett & Cavaye, but there introduces lack of fit problems without raising R-squared.) The remaining negative factors are dimensionality and professional practice assessment. The regression equation is now: $z = 0.796 - 0.284\text{prop} - 0.177\text{D} + 0.303\text{hi}$, where hi is 0 for a low quality study and 1 for a high quality one.

No significant combinations of factors were found to influence effect sizes.

Summary

The mean correlation over all the studies was 0.69, indicating definite evidence of agreement between peer and teacher marks on average. However, an r value of 0.69 indicates that less than half of the variation in peer marks is associated with variation in teacher marks. The mean effect size excluding the

unusual study is -0.02, not significantly different from 0. Even when the unusual study is included, the new mean of 0.24 is still not statistically significant. This also supports the conclusion that peer marks agree well with teachers' marks on average.

Analyses have identified the following variables as likely to be influential in terms of improving agreement between faculty and peer assessments. The variables are shown in decreasing order of significance as identified in this study.

- Peer assessments which require marking of several individual dimensions appear to be less valid than peer assessment which requires a global judgement based on well understood criteria. The optimum approach may be to require an overall judgement but entailing consideration of several dimensions or criteria.
- The nature of the assessment task appears to influence the validity of peer assessments. Peer assessment of academic products and processes seems to correspond more closely to faculty ratings than peer assessment in the context of professional practice.
- Studies that are well designed appear to give rise to better peer-teacher agreements than those with poor experimental designs.
- There is no evidence to support the superiority of multiple peer ratings over ratings by singletons. Ratings by very large numbers of peers (20+) appear to lead to poorer agreement.
- There are no clear differences in validity of peer assessments in terms of the subject area in which they take place, but peers in medically related subjects have a tendency to agree less well in some cases.
- Student familiarity with, and ownership of, criteria tends to enhance peer assessment validity.

Analyses also suggest that

- Peer assessment carried out on advanced level courses is no more valid than that conducted on introductory courses, in general.

The combination of a high quality study, an academic task, and a global judgement based on consideration of several dimensions or criteria would appear to lead to the highest correlation between peers and faculty.

The smallest effect sizes representing close peer-teacher agreement would appear to arise from judgements based on consideration of several dimensions or criteria, and from using less than 20 peers.

Discussion

Many predictions have been supported by the results of data analyses while some have not. Assessment using many individual dimensions seems more difficult than assessment using global judgements or with few dimensions. Assessment of a number of dimensions separately may involve marking each out of a sub-total. For example, a particular dimension or criterion may be awarded a maximum of 5 or 10 points. Thus, each division within the range can represent up to 20% of the total marks for that criterion. Small "errors" in each may add up to a large error overall. In addition, students may be reluctant to use extremely high or low ratings, which can amplify such problems.

The analysis found that peer assessment in the area of professional practice corresponded less well with faculty assessments than the marking of academic

products and processes. This may be explained by greater student familiarity with academic products and academic processes they have experienced for much of their formal education, than with professional practice which requires them to learn a new set of skills.

High quality studies appear to be associated with better peer-faculty agreement than studies of lower quality. If the studies rated as low quality also involved less-than-clear implementation, then it is understandable that students may have been confused about important elements of the exercise. For example, they may not have had complete understanding of the assessment mechanisms or of the criteria they were to use. Lack of understanding and confusion can readily lead to inaccurate marking.

It is surprising that there is little indication that peer assessment seems to be more valid in upper level courses in the present study, given that senior students are likely to have a better understanding of their discipline than their junior peers; are more likely to have a good understanding of the criteria by which they judge work; and may also have had previous experience of peer assessing. Perhaps the lack of differentiation between students in beginner and advanced courses in the present study indicates that participants in peer assessment studies are generally very well prepared for the task.

A very large number of assessors appears to produce marks that resemble those of the teacher less well than marks produced by a smaller number of raters or singletons. We were surprised to find that singletons performed as well as larger groups of students, given that it is generally acknowledged that multiple ratings are superior to single ones (e.g., Cox, 1967; Fagot, 1991). It has been argued that the use of multiple raters tends to improve reliability by increasing the ratio of true score variance to error variance (e.g., Ferguson, 1966). Some studies (e.g., Magin, 1993) have found that, while individual students may be poor judges, the reliability of averaged scores increases with the number of raters. However, it may be that when students work together in very large groups, some diffusion of responsibility occurs and marking becomes less thoughtful or careless. It is certainly the case that the degree of "social loafing" or "free-rider" effects within groups can become more pronounced as the size of a group increases (e.g., Kerr & Bruun, 1983). Kerr and Bruun found that, as group size increased, motivation of individuals decreased. Their experiments to investigate the "free-rider" effect suggested that task motivation of group members is sensitive to the perceived dispensability of their efforts for group success. Similarly, Latané, Williams, and Harkins (1979) found a sizeable decrease in individual effort when performing in groups compared with when performing alone, a phenomenon which they termed "social loafing." They attributed this effect to the difficulty of identifying individual contributions to a group effort.

Results from the present study may be compared with those from a similar meta-analysis of self-assessment studies (Falchikov & Boud, 1989) which suggested that well designed and reported studies were associated with closer correspondence between student and teacher marking than was the case in poorly designed ones. The self-assessment study also found that the level of the course of which the assessment was a part was another salient variable, with better agreement between faculty and students occurring in advanced level courses

than in lower level ones. Falchikov and Boud also found that, on the whole, student assessors in the area of science agreed more closely with teachers than students in other subject areas.

Thus, both studies find that well designed studies tend to produce better teacher-student agreements than poorly designed ones, and both note that more recent studies are better designed than older ones. Both suggest that when students are required to make judgements with little or no guidance, assessments are less accurate than when criteria are explicit and well understood. The present study, however, has found no clear subject area differences, whereas the earlier self assessment study found that self assessments in science were more accurate than those in any other subject discipline. We found no course-level differences in peer-teacher marking correspondence, whereas Falchikov and Boud found that level of course appeared to be a salient variable. These differences raise a question about possible differences between the acts of self and peer assessment. Self assessment is usually a private activity which may involve little or no knowledge of the work or performance of others. However, many of the peer assessment studies which make up the present corpus involve assessment of oral presentations or professional practice in a group context. Thus, the act of assessing takes place within a public domain where comparisons between performances become possible and ranking of peers becomes less difficult for students.

Some caution should be exercised in the interpretation of the results of the present study due to the presence of some very small sample sizes. In addition, there may be some liberal readings of the data due to the combination of variables in several ways. Nonetheless, the present study has increased our knowledge about this growing practice and provided some useful insights into the marking aspect of peer assessment which should be of practical use to both practitioners and educational researchers.

Recommendations to Practitioners for Implementing Peer Assessment

Based on the results of this meta-analysis, some recommendations for implementing peer assessment using marks or grades in higher education may be made. If the primary reason for introducing peer assessment is the degree of correspondence between peer and faculty marks, the following advice may be useful:

1. Avoid using very large numbers of peers per assessment group.
2. Conduct peer assessment studies in traditional academic settings and involve students in peer assessment of academic products and processes.
3. Do not expect student assessors to rate many individual dimensions. It is better to use an overall global mark with well understood criteria.
4. Involve your students in discussions about criteria.
5. Pay great attention to the design, implementation and reporting of your study.
6. Peer assessment can be successful in any discipline area and at any level.
7. Avoid the use of proportions of agreement between peers and teachers as a measure of validity.

It is also important to remember that peer assessment has many formative benefits in terms of improving student learning (e.g., Boud, 1988). Student peer assessment can successfully focus on the provision of feedback and may also be used in the absence of marking (e.g., Falchikov, 1995).

Future Work in This Area

Implementations of peer assessment are likely to continue. Increasing use of group project work and the frequent requirement to allocate individual marks for it are also likely to continue to be features of higher education. Peer assessment provides a way of achieving individual marks (e.g., Goldfinch & Raeside, 1990). Research in this area is well advanced and Jarvis and Quick (1995) argue that the mathematics of the process should be such that "ensuring a good team performance brings more reward than trying to outdo one's team-mates at the expense of team performance" (p. 179), while at the same time, giving due recognition to good, and not-so-good, contributors. The formative benefits of peer assessment are not in doubt, and increasing class sizes and demands from future employers for students with generic skills, which may be improved by peer assessment, constitute practical reasons for its continuance. The present analysis has provided some useful information relating to what constitutes an optimal context for successful peer assessment.

There are a number of important areas which deserve further investigation. These include

- Further exploration of the interactions between variables should be carried out as soon as more studies are added to the present corpus.
- An investigation of the effects of repeated experience of peer assessment is another important question to investigate in future work. The present study was not able to investigate these effects, as such data as were available were dependent and were combined before entry into the present analysis.
- So far, no work on gender effects has been conducted. This issue, too, deserves the attention of researchers. Gender effects are present in a wide variety of social and academic situations, and it is possible that they may also play a part in peer assessment. A study by Falchikov and Magin (1997) provides a methodology to enable investigation of gender effects in peer assessments.
- Further investigation of reliability and bias in the context of the single-multiple marker issue is desirable, particularly given that the present multivariate analysis was able to investigate only the comparison between very large groupings and all others.
- Investigations into friendship (or enmity) effects and their potential for bias might also be carried out.

Acknowledgements

Thanks are due to Mary McCulloch, Emma Forster and Jan McArthur for research assistance in the preparation of this paper, to Mairi Anderson and Dorothy Millar of the Dunning Library, Napier University, and to David Boud, Greg Michaelson, Gillian Raab and anonymous reviewers for very helpful comments on an earlier draft.

References

Bangert-Drowns, R. L., Wells-Parker, E., & Chevillard, I. (1997) Chapter 12 Assessing the methodological quality of research in narrative reviews and meta-analyses. In Bryant, K. J., Windle, M., & West, S. G. (Eds.), *The science of prevention. Methodological advances from alcohol and substance abuse.* Washington, DC: American Psychological Association.

Begg, C. B. (1994). Chapter 25 Publication bias (pp. 339- 422). In Cooper, H. and Hedges, L. V. (Eds.) *The Handbook of Research Synthesis.* New York: Russell Sage Foundation.

Billington, H. L. (1997). Poster presentations and peer assessment: novel forms of evaluation and assessment. *Journal of Biological Education, 31*(3), 218-220.

Boud, D. (Ed.). (1988) Developing student autonomy in learning (2nd ed.). London: Kogan Page.

Boud, D. J., & Tyree, A. L. (1979). Self and peer assessment in professional education: A preliminary study in law. *Journal of the Society of Public Teachers of Law, 15*(1) 65-74.

Burke, R. J. (1969). Some preliminary data on the use of self-evaluations and peer-ratings in assigning university course grades. *Journal of Educational Research, 62*(10), 444-448.

Burnett, W., & Cavaye, G. (1980). Peer assessment by fifth year students of surgery. *Assessment in Higher Education, 5*(3), 273-278.

Butcher, A. C., Stefani, L. A. J., & Tariq, V. N. (1995). Analysis of peer-, self- and staff-assessment in group project work. *Assessment in Education, 2*(2), 165-185.

Catterall, M. (1995). Peer learning research in marketing. In *Enhancing student learning through peer tutoring in higher education,* (pp 54-62). Jordanstown: University of Ulster, Educational Development Unit.

Chatterji, S., & Mukerjee, M. (1983). Accuracy of self assessment and its relation with some psychological and biographical factors. *New Zealand Journal of Psychology, 12,* 28-35.

Cooper, H. (1998) *Synthesizing research, A guide for literature reviews* (3rd ed.). Thousand Oaks, CA: Sage Publications.

Cox, R. (1967). Examinations and higher education: A survey of the literature. *University Quarterly,* 292-340.

Cross, K. P. (1981). *Adults as Learners.* San Francisco: Jossey-Bass.

D'Augelli, A. R. (1973). The assessment of interpersonal skills: a comparison of observer, peer and self ratings. *Journal of Community Psychology, 1,* 177-179.

Denehy, G. E., & Fuller, J. L. (1974). Student peer evaluation: an adjunct to preclinical laboratory evaluation. *Journal of Dental Education, 38*(4), 200-203.

Eisenberg, T. (1965). Are doctoral comprehensive examinations necessary? *American Psychologist, XX,* 168-169.

Ewers, T., & Searby, M. (1997). Peer assessment in music. *The New Academic, 6*(2), 5-7.

Eysenck, H. J. (1978). An exercise in mega-silliness. *American Psychologist, 33,* 517.

Fagot, R. (1991). Reliability of ratings for multiple judges: intraclass correlation and metric scales. *Applied Psychological Measurement, 15*(1), pp. 1-11.

Falchikov, N. (1999). Encouraging student autonomy and cooperative learning (in preparation), Napier University, Edinburgh.

Falchikov, N. (1995). Peer Feedback Marking: developing peer assessment. *Innovations in Education and Training International, 32*(2), 175-187.

Falchikov, N. (1994). Learning from peer feedback marking: student and teacher perspectives. In H. C. Foot, C. J. Howe, A. Anderson, A. K. Tolmie, & D. A. Warden (Eds.), *Group and interactive learning* (pp. 411-416). Southampton and Boston: Computational Mechanics Publications:

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

Falchikov, N., & Boud, D. (1989). Student Self-Assessment in Higher Education: a meta-analysis. *Review of Educational Research, 59*(4), 395-430.

Falchikov, N., & Magin, D. (1997). Detecting gender bias in peer marking of students' group process work. *Assessment and Evaluation in Higher Education, 22*(4), 393-404.

Ferguson, G. A. (1966). *Statistical analysis in psychology and education* (2nd ed.). New York: McGraw Hill.

Fineman, S. (1981). Reflections on peer teaching and peer assessment - an undergraduate experience. *Assessment and Evaluation in Higher Education, 6*(1), 82-93.

Freeman, M. (1995) Peer assessment by groups of group work, *Assessment and Evaluation in Higher Education, 20*(3), 289-300

Frieson, D. D., & Dunning, G. B. (1973). Peer evaluation and practicum supervision. *Counselor Education and Supervision, 12*, 229-235.

Fry, S. A. (1990). Implementation and evaluation of peer marking in higher education. *Assessment and Evaluation in Higher Education, 15*(3), 177-189.

Fuqua, D. R., Johnson, A. W., Newman, J. L., Anderson, M. W., & Gade, E. M. (1984). Variability across sources of performance ratings. *Journal of Counselling Psychology, 31*(2) 249-252.

Glass, G. V., & Smith, M. L. (1978). Reply to Eysenck. *American Psychologist, 33*, 517-518.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research.* Beverly Hills, CA: Sage.

Gray, T. G. F. (1987). An exercise in improving the potential of exams for learning. *European Journal of Engineering Education*, 12(4), 311-323.

Goldfinch, J., & Raeside, R. (1990). Development of a peer assessment technique for obtaining individual marks on a group project. *Assessment and Evaluation in Higher Education*, 15 (3), 210-225.

Guilford, J. D. (1965). *Fundamental statistics in psychology and education* (4th ed.) New York: McGraw Hill.

Hammond, K. R., & Kern, F., Jn. with Crow, W. J., Githers, J. H., Groesbeck, B., Gyr, J. W., & Saunders, L. H. (1959). *Teaching Comprehensive Medical Care.* Cambridge, MA: Harvard University Press.

Harris, M. M., & Schaubroeck, J. (1988). A meta-analysis of self-supervisor, self-peer and peer-supervisor ratings. *Personnel Psychology, 41*, 43-62.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* San Diego: Academic Press, Inc., Harcourt Brace Jovanovich, Publishers.

Hembree, R. (1988). Correlates, causes, effects and treatment of test anxiety. *Review of Educational Research, 58*(1), 47-77.

Hughes, I. E., & Large, B. J. (1993). Staff and peer-group assessment of oral communication skills. *Studies in Higher Education, 18*(3), 379-385.

Hunter, D., & Russ, M. (1996). Peer assessment in performance studies. *British Journal of Music Education, 13*, 67-78.

Hunter, J., Schmidt, F., & Jackson, G. (1982). *Meta-analysis: cumulating research findings across studies.* Beverly Hills, CA: Sage.

Jacobs, R. M., Briggs, D. H., & Whitney, D. R. (1975). Continuous-progress education: III. Student self-evaluation and peer evaluation. *Journal of Dental Education, 39*(8), 535-541.

Jarvis, P., & Quick, N. (1995). Innovation in engineering education: the "PAMS" project. *Studies in Higher Education, 20*(2), 173-185.

Kaimann, R. A. (1974). The coincidence of student evaluation by professor and peer group using rank correlation. *The Journal of Educational Research, 68*(4), 152-153.

Kegel-Flom, P. (1975). Predicting supervisor, peer, and self- ratings of intern performance. *Journal of Medical Education, 50,* 812-815.

Kerr, N. L., & Bruun, S. E. (1983) Dispensability of member effort and group motivation losses: free-rider effects. *Journal of Personality and Social Psychology, 44*(1), 78-94.

Kelmar, J. (1992, February). Peer assessment: a study of graduate students, Paper presented at the *Forum on Higher Education Teaching and Learning - the Challenge* conference, The Teaching and Learning Group, Curtin University of Technology, Perth, WA.

Korman, M., & Stubblefield, R. L. (1971). Medical school evaluation and internship performance. *Journal of Medical Education*, 46, 670-673.

Kwan, K-P, & Leung, R. (1996). Tutor versus peer group assessment of student performance in a simulation training exercise. *Assessment and Evaluation in Higher Education, 21*(3), 205-214.

Latané, B., Williams, K., & Harkins, S. (1979). Many hands make light work: the causes and consequences of social loafing. *Journal of Personality and Social Psychology, 37*(6), 822-832.

Lennon, S. (1995). Correlations between tutor, peer and self assessments of second year physiotherapy students in movement studies. In *Enhancing student learning through peer tutoring in higher education* (pp. 66-71). Jordanstown: University of Ulster, Educational Development Unit.

Linn, B. S., Arostegui, M., & Zeppa, R. (1975). Performance rating scale for peer and self assessments. *British Journal of Medical Education 9*, 98-101.

Magin, D. (1993). Should student peer ratings be used as part of summative assessment? *Higher Education Research and Development, 16,* 537-542.

Magin, D. J., & Churches, A. E. (1988). What do students learn from self and peer assessment? Proceedings of the Australian Society for Educational Technology, Canberra, 27-29 September, 224-233.

Marcoulides, G. A., & Simkin, M. G. (1991). Evaluating student papers: the case for peer review, *Journal of Education for Business, 67,* November/December, 80-83.

Melvin, K. B., & Lord, A. T. (1995). The Prof/Peer method of evaluating class participation: interdisciplinary generality. *College Student Journal, 29*, 258-263.

Montgomery, B. M. (1986). An interactionist analysis of small group peer assessment. *Small Group Behavior, 17*(1), 19-37.

Morton, J. B., & Macbeth, W. A. A. G. (1977). Correlations between staff, peer, and self assessments of fourth-year students in surgery, *Medical Education, 11*(3) 167-170

Mowl, G., & Pain, R. (1995) Using self and peer assessment to improve students' essay writing: a case study from geography. *IETI, 32*(4), 324-335.

Newstead, S., & Dennis, I. (1990). Blind marking and sex bias in student assessment. *Assessment and Evaluation in Higher Education, 15*, 132-139.

Newstead, S., & Dennis, I. (1994). Examiners examined: the reliability of exam marking in psychology. *The Psychologist, 7*(5), 216-219.

Ngu, A. H. H., Shepherd, J., & Magin, D. (1995). Engineering the "Peers" system: the development of a computer-assisted approach to peer assessment. *Research and Development in Higher Education, 18*, 582-587.

Oldfield, K. A., & Macalpine, M. K. (1995). Peer and self-assessment at tertiary level - an experimental report. *Assessment and Evaluation in Higher Education, 20*(1), 125-131.

Orpen, C. (1982). Student versus lecturer assessment of learning: a research note. *Higher Education, 11,* 567-572.

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*(3), 239-250.

Pease, D. (1975). Comparing faculty and school supervisor ratings for education students. *College Student Journal, 9*(1), 91-94.

Piaget, J. (1971, written 1969 and translated by Derek Coltman). *Science of education and the psychology of the child.* Longman: London.

Ritter, L. (1997). An educreational approach to the teaching of history in an Australian College of Advanced Education. In P. Ritter, *Educreation and Feedback: Education for creation, growth and change* (pp. 391-410). Oxford: Pergamon Press.

Rushton, C., Ramsey, P., & Rada, R. (1993). Peer assessment in a collaborative hypermedia environment: a case study. *Journal of Computer-Based Instruction, 20*(3), 73-80.

Shadish, W. R., & Haddock, C. K. (1994). Combining estimates of effect size. In H. Cooper & L. V. Hedges (Eds.), *The Handbook of Research Synthesis* (pp. 261-281). New York: Russell Sage Foundation.

Stefani, L. (1992). Comparison of collaborative, self, peer and tutor assessment in a biochemical practical. *Biochemical Education*, 20(3), 148-151.

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

Topping, K. J. (1998). Peer assessment between students in college and university. *Review of Educational Research, 68*(3), 249-276.

Vygotsky, L. S. (1962). *Thought and Language.* Cambridge, MA: MIT Press.

Wiggins, N., & Blackburn, M. (1969). Prediction of first-year graduate success in psychology: peer ratings. *The Journal of Educational Research, 68*(2), 81-85.

Authors

NANCY FALCHIKOV was, until very recently, a senoir lecturer in psychology in the Department of Psychology and Sociology at Napier University. She is now attached to the Educational Developments Service at Napier University, Bevar House; n.falchikov@napier.ac.uk. Her research interests include ways of improving student learning and involving students in assessment and peer tutoring in higher education.

JUDY GOLDFINCH is a senior lecturer in the Department of Mathematics at Napier University, Sighthill Court, Edinburgh, UK, EN11 4BN; j.goldfinch@napier.ac.uk. Her research interests include peer assessment aftrer group activities, as well as other forms of assessment, particularly computer-aided assessment.