

Rust, C, Price, M and O'Donovan, B

Improving students' learning by developing their understanding of assessment criteria and processes.

Rust, C, Price, M and O'Donovan, B (2003) Improving students' learning by developing their understanding of assessment criteria and processes. *Assessment and Evaluation in Higher Education*, 28 (2). pp. 147-164. ISSN 0260-2938

Doi: 10.1080/0260293032000045509

This version is available: <http://radar.brookes.ac.uk/radar/items/0ca6b67d-5ee4-8e2a-17b1-8ffd8791c26d/1/>

Available in the RADAR: October 2009

Copyright © and Moral Rights are retained by the author(s) and/ or other copyright owners. A copy can be downloaded for personal non-commercial research or study, without prior permission or charge. This item cannot be reproduced or quoted extensively from without first obtaining permission in writing from the copyright holder(s). The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the copyright holders.

This document is the preprint version of the journal article. Some differences between the published version and this version may remain and you are advised to consult the published version if you wish to cite from it.

Title:

Improving students' learning by developing their understanding of assessment criteria and processes

Chris Rust, Margaret Price and Berry O'Donovan

Oxford Brookes University

Abstract:

This paper reports the findings of a two-year research project focussed on developing students' understanding of assessment criteria and the assessment process through a structured intervention involving both tacit and explicit knowledge transfer methods. The nature of the intervention is explained in detail, and the outcomes are analysed and discussed. The conclusions drawn from the evidence are that student learning can be improved significantly through such an intervention, and that this improvement may last over time and be transferable, at least within similar contexts. This work is a development within a longer and on-going research project into criterion-referenced assessment tools and processes which has been undertaken in the pursuit of a conceptually sound and functional assessment framework that would promote and encourage common standards of assessment; that project is also summarised.

INTRODUCTION

Within Higher Education there is an increasing acceptance of the need for a greater transparency in assessment processes, and moves have been made to make methods of assessment clearer to all participants. This paper is concerned with the extent to which students understand these processes and how we might improve their understanding of them. It presents the development and planning of a two year project involving the transfer of knowledge of the assessment process and requirements to students in a variety of ways, in particular, through a structured process involving both tacit and explicit knowledge transfer methods. The aims of this project were to improve the students' performance and promote independent learning through enhancing their ability to assess the work of others and, in consequence, their own work against given marking criteria. The initial findings of the first year of the project, the methodology and its background were first reported at the 8th Improving Student Learning Symposium in Manchester, England, and first published in the conference proceedings (Price, O'Donovan and Rust, 2001). The success of the project, and a replication of the exercise with a second cohort the following year, has now been evaluated from a number of perspectives the most important of which being by gauging the subsequent effect on the students' performance. A further evaluation of the longer-term effect on performance has also been carried out on the first cohort.

BACKGROUND

This work is a development within an on-going research project into criterion-referenced assessment tools and processes which has been undertaken in the pursuit of a conceptually sound and functional assessment framework that would promote and encourage common standards of assessment. The earlier findings from this larger project have informed the development of this research and have already been reported elsewhere (Price and Rust, 1999; O'Donovan, Price and Rust, 2001) and are summarised below.

CONTEXT

The research project into criterion-referenced assessment tools and processes commenced in 1997 against a background of growing national concern in the UK about marking reliability, standards and calls for public accountability (Laming, 1990; Newstead and Dennis, 1994). At a national level within the UK compelling pressure was beginning to be applied to higher education institutions to maintain high academic standards (Lucas and Webster, 1998). This pressure has been exacerbated over the last few years by an apparent fall in standards suggested by the rise from 25% to 50% in the proportion of good degree results (upper second class and first class degrees). This trend has been compounded by the rapid expansion of student numbers and a drastic cut in the unit of resource for UK higher education. The debate about standards was further informed by a national discussion on generic level descriptors (Otter, 1992; Greatorex, 1994; Moon, 1995; HEQC, 1996) which were

seen by some as a means of establishing common standards. The focus of this discussion tended to be on the need for explicitness, with the implication that if this was achieved it would be enough. Little, if any, mention was made about involving students in the process.

In response to this, the Quality Assurance Agency (QAA) embarked on a new quality assurance system, with three distinct elements – benchmarking, programme specifications, and a national qualifications framework – all intended to bring about the establishment of explicit degree standards. However it is interesting to note that when the benchmarks were published in May 2000 they were re-titled benchmarking statements. Arguably, this change recognised the failure of the process to clearly define explicit standards for all subjects. At a conference on Benchmarking Academic Standards (Quality Assurance Agency, 17th May 2000), Chairs of the QAA subject panels commented on the difficulties of defining threshold standards and using language which meaningfully conveyed level. However the benefit realised by the academic community from the *process* of drawing up the statements was emphasised. Prof. Howard Newby stated:

‘I would certainly want to assert the value to self-understanding in disciplines of debating the basis on which the discipline is conducted and what the students need in order to be able to participate in the community of scholars who practise it’
(QAA, Benchmarking Academic Standards Conference, 17th May 2000).

FIRST STEPS

The initial impetus to address the issues in this project came from an external examiner for the Business Studies programme at Oxford Brookes University who was a strong proponent of criterion-referenced assessment as a means of ensuring consistent standards between markers. Another external examiner was concerned to ensure common standards between modules. As a consequence of this a common criteria assessment grid was developed for the Business School and first piloted in the academic year 1997-98. The grid has 35 criteria plotted in matrix format against grades resulting in 'grade definitions' detailing acceptable performance for each criterion at each grade (one page of the grid detailing six criteria is reproduced in Figure 1). Staff select appropriate criteria for any given assessment to create a 'mini-grid' (see Figure 2 for an example). The main intention was to develop a comprehensive marking criteria grid to help establish common standards of marking and grading for Advanced Level undergraduate modules (those normally taken by second and third year students) across the Business programme enabling consistency in marking and easier moderation. Furthermore, it was hoped that the grid would have the additional benefits of providing more explicit guidance to students (resulting subsequently in better work), and making it easier to give effective feedback to the students.

[Type setting note - put in figures 1 and 2 here]

STAFF AND STUDENTS' VIEWS

The use of the grid has been evaluated through the views of staff and students as well as noting the feedback from external examiners.

The main conclusion of the initial paper was that, at least in its present form and usage, the grid failed to establish a common level - different tutors having taken the grid and used exactly the same grade definitions for a basic module (one normally taken by first year students) and an MBA module apparently without any difficulty. However, the paper further concludes that the findings had demonstrated that the use of such a grid could provide other real benefits. It could help to raise the quality of marking through greater consistency in marking both for a team of markers and for an individual marker, but this was more likely to be the case if the tutors had discussed the grid together before using it. It could also help provide, from the tutor perspective, more explicit guidance to students and thus potentially improve the quality of their work, although it appeared that this was only likely to be true for the most motivated students unless time was spent by tutors discussing with students the meaning of the criteria terms and grade definitions. Using the grid could also raise the quality of feedback to students and assist in focusing the marker's comments.

The initial mixed findings reflected many of the issues associated with criterion-referencing in the marking of more qualitative and open form assessment. Whilst many would agree that criterion-referenced assessment appeals to our notion of equity and fairness, it is not without its pitfalls, not least of which is the potential for

multiple interpretations of each criterion and grade definition by both individual staff members (Webster et al, 2000) and students.

The views of students were sought when they had experienced the grid on a variety of modules and more detailed findings have been reported elsewhere (O'Donovan et al, 2001). The students felt the criterion-referenced grid to be a well-conceived assessment tool and clearly recognised the potential of the grid and what it was trying to achieve. However it was also seen as of limited practical use if presented in isolation without the benefit of explanation, exemplars and the opportunity for discussion. The need for such aids resulted from the identification of several issues undermining the easy application of the grid. These included clarification of the meaning of terms and phrases; subjectivity and multiple interpretations of criteria and standards; a lack of match between published criteria and the feedback received, arguably, suggesting the presence of a 'hidden curriculum' (Sambell and McDowell, 1998; Webster et al, 2000).

IMPLICATIONS

A common view of both staff and students was the need for discussion to support the use of the grid, between staff in the marking process, and between staff and students to enhance students' understanding of the assessment process and as a result to improve their performance. Students also identified exemplars and further explanation as useful in making the assessment criteria more comprehensible.

When QAA experts are unable to make standards explicit after months of learned debate, arguably, we should, perhaps, begin to question the single-minded focus on explicit articulation of standards. It is difficult to relinquish the notion that academic standards can be documented and codified in such a way that they may be available for the passive consumption of all stakeholders in higher education. However, our research experience has been that, without active involvement through discussion and debate, the development of a common view on standards and level is problematic, if not impossible, even within a close knit community of a single academic department. Obstacles to the transfer of knowledge about standards and assessment requirements are accentuated when such knowledge requires transference to more 'novice' students undertaking modular courses in which they have very limited time to construct cohesive, 'objective', interpretations of assessment requirements. Obstacles yet further heightened in a broad-ranging, multidisciplinary and discursive subject such as business and management. A discipline in which many open form assessments, i.e. essays and reports, are set on topics which are integrative in nature.

Consequently, we suggest that the imprecision inherent in passively presented verbal description requires that consideration be given to other ways of achieving shared understanding of criteria and standards. Arguably, in its present quasi-scientific form the grid incorporates too great an assumption about the nature of knowledge to be transferred and our ability to transfer it.

TRANSFERRING TACIT KNOWLEDGE

We conjecture knowledge of assessment criteria and process is a combination of tacit and explicit knowledge. Accordingly, careful articulation, although ideal for the transfer of explicit knowledge, is not, in itself, sufficient to share knowledge of the assessment process. It seems that the ‘missing’ information transferred through the process of discussion and further explanation, imitation or practice is carried in the minds of the assessors (initially in the mind of the assignment writer). Such knowledge can be described as tacit (tacit knowledge in this context being defined as something that we know but we find impossible or, at least, extremely difficult to express). Experts on knowledge management (see for instance Nonaka, 1992; Baumard 1999) suggest that tacit knowledge transference takes place through the sharing of experience - socialisation processes involving observation, imitation, practice and dialogue. So discussion and shared experiences of marking and moderation over time among staff enables the sharing of tacit knowledge resulting in more standardised marking. It follows that inviting students into this shared experience should also enable more effective knowledge transfer of assessment processes.

However, mirroring the complexity of the assessment process, the means of enhancing understanding of assessment does not lie in abandoning explicit knowledge transfer methods completely for those geared to the transfer of tacit knowledge. Whilst Baumard convincingly argues that the tacit dimension can be crushed or stubbed out by an over emphasis on explicit knowledge (1999, p.194), plausibly, the

transfer of knowledge on assessment requirements lies within a combination of explicit and tacit knowledge transfer methods that are mutually complementary and interdependent. For example, students identified the use of exemplars as beneficial but these do not fit clearly within either definition of explicit or tacit knowledge (or the knowledge transfer processes of articulation or socialisation). In considering a number of different practices used in assessment a continuum of different methods lying between explicit and tacit knowledge transfer can possibly be identified (from explicit descriptors at one end to the use of exemplars, and self and peer assessment at the other). All these methods have been used to a greater or lesser extent to aid student learning, but their combination may be the key to enhancing student performance through an understanding of the assessment process.

So although the research to date suggests that the grid is inadequate in isolation it may still have a role in a framework of assessment processes encompassing a multi-faceted approach which derives from such a continuum.

THE RESEARCH PROJECT

Following on from the findings that the grid was not, in itself, capable of being level-specific, and the students' request for an introduction to the grid in their first year of undergraduate studies, the following two year project was devised for a large (300+) basic (first year) undergraduate business module.

Taking methods from along the explicit/tacit knowledge transfer continuum a structured programme was devised to engage the students in a series of activities that combined the articulation of marking criteria, the use of exemplars, the application of marking criteria, dialogue and self-assessment.

We were also influenced in the design of this pilot by claims from the United States (Nelson, 1994) of the potential effectiveness of student marking exercises, along with studies from the UK which have shown significant improvement in the work of students involved in marking using model answers (Forbes and Spence, 1991; Hughes, 1995, et al).

PROCESS

The intervention took place in the final three weeks of the students' first term on the degree programme. It involved students in preparation work, attendance at a workshop, and the submission of a self-assessment sheet along with their coursework to be handed in at the end of the first term (three weeks after the workshop).

The detail of this process was as follows:

- a) A week before the workshop all students on the module were provided with 2 sample assignments (one excellent piece of work and one a borderline pass) and marksheets including assessment criteria and grade definitions. Students were asked to individually complete a marksheet providing a grade, marks and

rationale/feedback for each of the assignments before coming to the workshops.

- b) Workshops (90 minutes long) were held for all students in groups of 40. These were held within the module's time frame but were in addition to the weekly lecture and seminar. The workshops were structured in the following way:
- i. Small group discussion of initial marking of sample work;
 - ii. Feedback of small groups' agreed grades and rationale to plenary
 - iii. Tutor-led comparison of provided rationales with criteria;
 - iv. Tutor explanation of each criterion;
 - v. Small groups review assessment and grade;
 - vi. Final small group report to plenary of grade for each piece of work;
 - vii. Tutor provided annotated and marked versions of samples and discussed tutor assessment and mark.

The small group discussions allowed the student to compare and justify their initial assessment of the work against that of others as well as allowing the declared grade to be the responsibility of the small group. However the students were asked explicitly not to change their initial grading on their individual sheets.

- c) Three weeks later, students submit their coursework along with a completed self-assessment sheet.

The feedback sheets for the sample assignments and for the self-assessment were the same, incorporating comments, an assessment grid, a grade and a mark

DESIGN FRAMEWORK FOR QUANTITATIVE ANALYSIS

The intervention was designed to run twice, in successive years, with two different cohorts on exactly the same module and in exactly the same way, so that it would provide baseline and treatment comparisons for statistical analysis, and also a transfer comparison in the second year, as detailed below:

- i) A baseline comparison - comparison of the assessment performance of the group of attendees at the assessment workshop with the non-attending group on a module (7508) taken by both groups before the training was carried out. This comparison to be carried out in two successive years (1999 and 2000) with successive cohorts.
- ii) A treatment comparison - comparison of the assessment of the group of attendees at the assessment workshop with the non-attending group on a module (Module 7009) taken by both groups within which the workshop was carried out before the assessment. This comparison also to be carried out in two successive years (1999 and 2000) with the same successive cohorts as in (i).

- iii) A transfer comparison - comparison of the group of attendees at an assessment workshop with the non-attending group on a module (Module 7026) taken by both groups one academic year later than Module 7009 (in 2001)

When the project was completed, between-groups comparisons were carried out to examine the effects of Attending vs. Not Attending on the baseline module (7508 x 2 years), the treatment module (7009 x 2 years) and the transfer module (7026 x 1 year). Data preparation was carried out in Microsoft Excel. The between-group comparisons used independent group t-tests and were carried out in *AlStats*. Because of the number of t-tests used ($n = 5$) a Bonferroni correction was applied to convert the 5% significance level (one-tailed) to the 1% level (one-tailed). In advance of the study a significance criterion of 0.01 (1%) was set. The effects ratio (mean difference divided by standard deviation) was also calculated for each comparison with 0.5 considered a moderate effect (Cohen, 1969).

COLLECTING ADDITIONAL DATA

The outcome of the project was further monitored using both qualitative student and staff evaluation, and quantitative data on the students' subsequent self-assessment.

Establishing the extent of student initial understanding of the assessment criteria and process, i.e. prior to the process

In order to test out the students' initial understanding it was necessary to test their application of the criteria and notion of standards. Through the use of the sample assignments the students' initial attempts at marking were collected via the completed marksheets from those students who participated in the workshop. The marksheets provided the grade (A-C or F) and mark (%) awarded, reasons for the decision/identified strengths and weaknesses of the work and a completed assessment grid showing the application of the marking criteria and grade definitions in marking.

At the workshop the student contribution was monitored through non-participant observation. Notes were taken on the grades awarded by the students for the sample assignments following small group discussion as well as the rationales provided for those grades. This generally provided more detailed reasons for the award of grades than the marksheets and a clearer indication of the level of understanding of each criterion being applied.

Monitoring the process

Non-participant observation during the workshops was also used to monitor the response of the students to the socialisation process. Questions asked by the students and discussions during explanation of the criteria were noted as well as the effects of the reapplication of the criteria following further explanation.

Using a questionnaire with position statements and Likert scales of agreement/disagreement, students were asked to evaluate the workshop in terms of its

effectiveness on their understanding of criteria and assessment processes and their levels of confidence in completing their assessed work and applying the criteria to their own work.

Effects of the process

Student ability to self-assess

All students on the module were required to submit a self-assessment of their assignment with their assessed coursework. It was the intention of the researchers to compare the student's self-assessed grade with that of the marker to provide an indication of the students understanding of level. This was only in fact possible with the second year cohort (see below).

Staff perceptions

In the first year of the project, using informal interview/unstructured discussion, the seminar tutors (also the first markers of the assessed coursework) were asked about student response to the intervention and their perceptions of how well the students had done the assessed coursework. In particular, those who had taught on the module in the previous year were asked to comment on how the work compared. These perceptions were gathered in the knowledge that the influence of norm-referenced marking is the dominant model within higher education and 'naturally' preferred by most markers (Professor A. Wolf, SHRE Assessment Network Meeting, 13/3/2000).

Therefore the measurement of student performance alone may not be sufficient to establish the effect of the assessment process.

Methodological issues

Participant self-selection

The participating students under scrutiny were self-selected due to the non-compulsory nature of the pilot process, as such we were concerned that these student participants might not be representative of the cohort population. Results could be easily skewed with the participant sample differing from the population in key attributes such as ability or motivation.

The results were tested in two ways: firstly, the standard deviation of the marks for each group were calculated and compared to surface any anomalies. Secondly, the marks achieved by the participating students were compared against those of the non-participating students on a piece of coursework on another module submitted prior to the workshop sessions (a baseline comparison).

Data contamination from exemplars

Identifying the sample of students that had taken part in the full process of the pilot by attending the workshops was straightforward. However it must be recognised that the non-participant remainder of the cohort were not completely isolated from the

process. All students were issued with the original sample assignments for their initial preparatory assessment and, more importantly, it is inevitable that the tutor marked and annotated versions of the samples will have been widely circulated. So although many of the cohort did not take part in the knowledge transfer process within the workshop many may have had the benefits of exemplars of the standards. The effect of this may well have been to diminish any differences between the performance of the participating sample and the rest of the cohort. However, this serves to heighten the significance of the active involvement of the participating students in the difference that did emerge between the two groups.

Norm referencing in marking

Even though the application of criteria-referenced marking should supplant the application of norm-referenced marking it seems that it remains an influencing factor in marking whichever is used (Professor A. Wolf, SHRE Assessment Network Meeting, 13/3/2000). (Norm-referencing indicating when students' performance within a group are simply compared and ranked, as opposed to criterion-referenced marking where more explicit definitions are determined of what is required, either to pass or, in a more complex application, for specific grades.)

The difficulty posed by norm-referencing within quantitative analysis of assessment results is that if assessors mark comparatively to an average, say of 55%, then inter-module comparison and longitudinal comparison of results can be fairly meaningless. The numerical description of 55% simply becomes the portrait of average whilst the

‘real’ average performance of two different cohorts could be very different.

Consequently, numerical descriptions can confer delusions of accuracy. Qualitative data from the markers about their perceptions of the quality of the students’ work help counter this effect of norm-referencing.

Student self-assessment

In the first year of the project, because of what with hindsight can be seen to have been a very silly error in the methodology (ie the assessment sheets on which both the students self-assessed and then the tutor assessed were photocopied for use on this research project, prior to their return to the student with their work), it was not possible to tell which assessments had been made by the student and which by the tutor, making any analysis impossible. This error was rectified in the second year.

FINDINGS

THE EXTENT OF THE STUDENTS’ UNDERSTANDING OF ASSESSMENT CRITERIA AND PROCESSES

The students’ initial attempts at grading and the workshop discussion showed that they exhibited more confidence in applying explicit, visible criteria - structure, presentation and referencing. These criteria were used extensively as justification of the grade awarded on individual marksheets and by the small groups in the workshop.

Conversely, the students found difficulty in explicitly applying more ‘invisible’ criteria - analysis, evaluation, etc. There was a deep reluctance to use these invisible criteria at the start of the process. Even when they were mentioned in the justification of the marks their application was weak. Following explication of the criteria the students in the workshops then began to apply them cautiously but still found it difficult to use them to justify marks. Many students commented on how difficult they found the marking task and their fear at exposing a lack of ability to assess.

It is interesting to note that students graded more conservatively than tutors did, even after involvement in the series of structured activities. This was more noticeable at the higher level than the threshold pass. This conservatism may have been as a result of a lack of confidence in applying the criteria; an expectation of higher standards or an indication that they are not used to seeing the full range of marks used.

Understanding of level

In the first year, the evaluation of students’ ability to assess was based on an analysis of 116 marksheets and marking grids completed and handed in by 116 of the 151 students who attended the workshops.

The initial grades taken from the marksheets for each of the sample assignments showed that almost every student had correctly identified the excellent and poor piece of work relative to each other. However, for each piece of work there was a range of grades awarded.

Assignment One had been graded as an A by the tutors, while the profile of marks awarded by the students included:

A	B+	C
34 students (29%)	16 students (14%)	63 students (54%)

(3 students did not grade this assignment)

Assignment Two was graded as C by the tutors and the students' assessment showed slightly more alignment, with the profile of marks awarded by students including:

B+	B	C	F
1 student (.9%)	22 students (19%)	87 students (75%)	3 students (3%),

(3 students did not grade this assignment)

Following the workshop activities, where the review of the application of assessment was in small groups, 8 out of 39 groups shifted their grades towards the then unknown tutor grading. It should be noted that one might expect the more 'extreme' marks to be moderated out of the system by the group process and that several groups were in line with tutor grading in the first place. None of the groups shifted away from tutor grading.

Student perceptions

Feedback from students indicated that they viewed the workshop very positively.

They felt that the activities and discussion had contributed 'a lot' to their

understanding of marking criteria and their assignment. Also, generally, they felt more confident about their assignment preparation although a small minority stated they felt less confident after the workshop because although they better understood the level required to pass they were concerned about their ability to meet it. Many students requested that the workshop be scheduled earlier before they undertook any coursework on other modules.

Markers' perceptions

In the first year of the project, markers perceived the standard of student coursework on the module to be higher than that of previous years but that there were indications of convergence of style and structure. The scripts were not separated to be marked and the markers were not aware which of the students comprised the participating sample. The markers could only form an overall impression of standard rather than improvement of particular students' work.

THE EFFECT ON PERFORMANCE STANDARDS - RESULTS OF STATISTICAL ANALYSIS

COHORT 1

Participants: 151 Non-participants 143

Module	Attending	Not Att	df	p	Mean	Std	Effects
--------	-----------	---------	----	---	------	-----	---------

	(mean)	(mean)			difference	dev.	ratio
7508	58.6	57.8	191	0.56	0.7606	9.08	0.08
7009	59.78	54.12	292	0.00001	5.649	9.49	0.6
7026	57.91	51.3	182	0.004	6.604	11.91	0.55

COHORT 2

Participants: 152 Non-participants 169

Module	Attending	Not Att	df	p	Mean	Std	Effects
	(mean)	(mean)			difference	dev.	ratio
7508	58.67	55.75	133	0.06	2.92	9.32	0.31
7009	59.86	52.86	319	0.00001	6.9716	10.03	0.69

As can be seen from the table of results above, with both cohorts there was no significant difference between the attendee and non-attendee groups in the baseline comparison on Module 7508, where the assessment was submitted before the training took place. (And power analysis of these figures provides power estimates in both cases which show the design was almost certainly adequate to have identified differences should they have existed - Power (5%) = 93% and 76% respectively.)

After the training, there is a significant difference between the results of those attending the training and for those who did not, for both cohorts ($p < 0.01$) on the initial module (7009), and that significance can still be identified one year later for the first cohort ($p < 0.01$) (on module 7026) although the figures may suggest, not surprisingly, a minor reduction in the effect. This is also reflected in the effects

ratios, 0.6 and 0.69 in the module on which the training took place (7009), and 0.55 in the module one year later (7026).

THE RESULTS OF STUDENT SELF-ASSESSMENT AND COMPARISON OF THEIR ACCURACY

All students were asked to complete and submit a marking criteria grid (see Fig 4) grading their work for each of the five criteria as well as giving an overall grade, but in the first year the data was not usable (as has already been explained above). In the second year, however, 140 (92%) workshop attendees and 150 (89%) non-attendees complied with the request and gave themselves an overall self-assessment grade but only a somewhat disappointing 68 (45%) attendees and 38 (22%) non-attendees additionally graded themselves for each of the individual criteria. The students self-assessments were then compared with the grades given by their tutors, and a comparison made between the workshop attendees and non-attendees to see if the former were able to be more accurate in their self-assessment.

To make a straight comparison of the accuracy of the self-assessments, a simple numerical system was devised whereby zero indicated that student and tutor had put the same grade, one indicated a one-grade difference (plus one if the student's grade was higher, ie an overestimate; minus one if the student's grade was lower, ie an underestimate), two indicated a two grade difference, etc. (The range of possible grades being A, B+, B, C, F).

Comparison of overall grades

An initial comparison of the overall grades given by the students in the two cohorts was disappointing because there did not seem to be any great difference; in fact, with 54 (39%) attendees and 68 (45%) non-attendees accurately predicting their grade and 27 (19%) attendees and 30 (20%) non-attendees only over predicting by one grade it looked as if anything the non-attendees were more accurate.

Attendees

Overall grade difference	No of students (n=140)	%
+3	1	.7%
+2	11	7%
+1	27	19%
0	54	39%
-1	38	27%
-2	9	6%

Non-attendees

Overall grade difference	No of students (n=150)	%
+3	8	5%
+2	18	12%
+1	30	20%
0	68	45%

-1	22	15%
-2	4	2.5%

However, on closer examination it is interesting to note that if we compare those either overestimating by two or three grades, and those underestimating, there is a difference with less attendees greatly overestimating and far more underestimating, and we will return to discuss this difference later.

	Attendees (n=140)	non-attendees (n=150)
Overestimating (+2 or +3)	12 (8.5%)	26 (17%)
Underestimating (-1 or -2)	47 (34%)	26 (17%)

Comparison of individual criterion grades

While from the data above, it appears that the non-attendees have actually been more accurate it would be possible to hypothesise that they may have got the right result but for the wrong reason so a comparison was done between the workshop attendees and non-attendees who had self-assessed individual criterion grades for each of the criteria. This was first done by simply totalling every grade difference there was between each of their self-assessments and the tutor's, regardless of whether they were over or under estimates (in other words a +1 difference and a -1 difference totalled as 2; plusses and minuses were not allowed to cancel each other out).

For those who had accurately predicted their overall grade, a higher percentage of workshop attendees (23% cf 12%) had been totally accurate getting every single criterion grade right too, but the numbers are very small, and otherwise there did not seem to be any obvious difference at all.

Students with accurate overall grade self-assessment

Total of individual criterion grade differences	Attendees (n=26)	Non-attendees (n=17)
0	6 (23%)	2 (12%)
1	5 (19%)	1 (6%)
2	9 (35%)	8 (47%)
3	2 (7.5%)	1 (6%)
4	1 (4%)	3 (17.5%)
5	3 (11.5%)	2 (12%)

Over and under estimation

However, remembering the apparent slight difference between the two groups regarding over and under estimating their overall grades (already reported above), it

was then decided to do one further analysis of the individual criterion grade differences looking at whether the grades were over or under estimations.

Criterion 1: presentation

grade difference student cf tutor	Attendees (n=68)	Non-attendees (n=37)
+3	1 (1.5%)	2 (5.5%)
+2	3 (4.5%)	1 (3%)
+1	13 (19%)	14 (38%)
0	42 (62%)	17 (46%)
-1	7 (10%)	3 (8%)
-2	2 (3%)	0

Criterion 2: content/range

grade difference student cf tutor	Attendees (n=67)	Non-attendees (n=38)
+3	1 (1.5%)	1 (3%)
+2	1 (1.5%)	3 (8%)
+1	13 (19%)	11 (29%)
0	26 (39%)	19 (50%)
-1	23 (34%)	4 (10.5%)
-2	3 (4.5%)	0

Criterion 3: use of literature/evidence of reading

grade difference student cf tutor	Attendees (n=68)	Non-attendees (n=37)
+4	1 (1.5%)	0
+3	1 (1.5%)	0
+2	0	4 (11%)
+1	11 (16%)	10 (27%)
0	34 (50%)	14 (38%)
-1	16 (24%)	9 (24%)
-2	5 (7%)	0

Criterion 4: analysis

grade difference student cf tutor	Attendees (n=67)	Non-attendees (n=37)
+3	0	2 (5%)
+2	4 (6%)	4 (11%)
+1	12 (18%)	6 (16%)
0	29 (43%)	20 (54%)
-1	18 (27%)	0
-2	4 (6%)	0

Criterion 5: evaluation

grade difference	Attendees (n=68)	Non-attendees (n=38)
------------------	------------------	----------------------

student of tutor		
+3	0	1 (2.5%)
+2	3 (4.5%)	6 (16%)
+1	13 (19%)	13 (34%)
0	30 (44%)	16 (42%)
-1	15 (22%)	2 (5%)
-2	7 (10%)	0

Although these are quite small numbers, especially in the case of non-attendees, there is a distinct pattern in these results with a higher percentage of non-attendees overestimating the grade and a higher percentage of attendees underestimating the grade. And the criteria for which this is arguably most marked are criteria four and five (analysis and evaluation).

Gender

Because of the suggestion that has been made in some of the literature on self-assessment that women may be more likely to underestimate their true worth while males may be inclined to overestimate (Thomas, 1990; Gibbs, 1991), it was then decided to return to the overall grade assessments and to divide them by gender to see if this could be a factor in these results. (The numbers of students who had self-assessed against each criterion was so comparatively small that to further break down those results by gender was rejected.)

Overall grade diff.	Female attendees (n=83)	Male attendees (n=57)	Total attendees (n=140)	Female n/attendees (n=52)	Male n/attendees (n=98)	Total n/attendees (n=150)
+3	1 (1%)	0	1 (0.7%)	0	8 (8%)	8 (5%)
+2	7 (8%)	4 (7%)	11 (7%)	3 (6%)	15 (15%)	18 (12%)
+1	17 (20%)	10 (18%)	27 (19%)	12 (23%)	18 (18%)	30 (20%)
0	31 (37%)	23 (40%)	54 (39%)	24 (46%)	44 (45%)	68 (45%)
-1	23 (28%)	15 (26%)	38 (27%)	10 (19%)	12 (12%)	22 (15%)
-2	4 (5%)	5 (9%)	9 (6%)	3 (6%)	1 (1%)	4 (2.5%)

Amongst the attendees there appears to be no identifiable difference between the males and females, however with the non-attendees there does seem to be more male overconfidence, 48% of 29% overestimating and 23% of 6% overestimating by two or more grades.

Comparison of the two male groups shows male non-attendees more confident than attendees – 42% overestimating their grade of 24%, and 26% of 4% overestimating by two or more grades. Conversely, 13% male non-attendees graded their assignment lower than their tutor in comparison to 35% of male attendees.

Females showed less of a startling difference – 25% female non-attendees underestimated their work cf 33% of female attendees.

COMMENT

TRANSFERRING TACIT KNOWLEDGE

The move towards greater transparency in assessment processes has been founded on the articulation of standards, levels and criteria for assessment in written format. The provision of information in such a format was considered sufficient to increase the participants' understanding of the processes. In the light of this pilot project and the experience of the QAA in seeking to establish benchmark standards, such assumptions about the transfer of knowledge of assessment processes need to be questioned. The QAA's failure to establish (through subject benchmarking) explicit standards may lie with the assumption that all aspects of the standards could be articulated and made explicit. This does not fit comfortably with the application of standards through the use of the traditional assessment model that relies on a normative, connoisseur approach. A connoisseur approach undertaken by those that 'regard assessment as akin to wine tasting - a high level activity that requires continued practice but that is pretty much impenetrable to the non-cognoscenti' (Webster et al, 2000, p. 73) . Such an approach appears to rely on a relationship between student and tutor developed over time to achieve the transfer of knowledge, both explicit and tacit, from novice to expert (Eraut, 1994). That transfer process

takes place gradually, if at all, through a complex social process relying on feedback and discussion. And even if one were to concede that in the past it may have been reasonably successful with a majority of students (and some may wish to challenge that), it is surely questionable whether in these times of increased student numbers, an increasingly diverse and 'part-time' student body, and diminished resources all leading to reduced staff-student contact that such a process can be relied upon to take place automatically - and certainly not for all students.

The findings from this project also point to the significant factor in knowledge transfer and sharing being the socialisation processes focused on in the workshop. Given that all students were provided with the samples of work prior to the workshop and annotated versions, given out at the workshop, were widely circulated among the whole student group, the workshop remains the distinguishing aspect of the process. Those students taking a full part in all the activities were seen to perform to a better standard than the rest of the cohort.

TRANSFER METHODS

Although the research demonstrates that students benefited from the complete process and they felt more confident about undertaking their coursework, it is not clear which aspect of the socialisation process had the greatest effect. Only those students who had practised marking the sample work, discussed their assessment and had further explication of the criteria demonstrated better than average performance; and

evidence from the literature on peer-marking using model answers (Forbes and Spence, 1991; Hughes, 1995) would suggest that it is being engaged with the process of marking as well as seeing examples of other work that significantly contributes to the student's subsequent improvement in performance. Certainly feedback from the students suggested that the use of examples contributed considerably to students' contextual understanding of assessment criteria and standards. The two sample pieces of coursework arguably provided vehicles through which the students could see the application of abstract criteria. Some students in the rest of the cohort may have benefited from consideration of the annotated exemplars that circulated following the workshop. Although there is no clear evidence of improved performance of these students the markers did feel the overall standard of work had improved. It can be acknowledged that tacit knowledge transfer is necessary for full understanding but sufficient understanding may be gained from activities that are found only part way along the explicit / tacit continuum. Further research may reveal the relative effectiveness of tacit knowledge transfer methods.

TRANSFERABILITY

Albeit that the evidence is based on only one cohort so far, there is encouraging evidence that a relatively simple intervention, taking a relatively small amount of course time can have an effect which can last over time and be transferred.

Admittedly the follow-up module was deliberately chosen because the nature of the assessment task, and the assessment criteria used, were similar so no grand claims for transferability can be made on this evidence but it would be interesting to continue

this research project to look at the performance of these students on very different types of assessment.

SELF-ASSESSMENT ACCURACY

Although disappointing in some ways that the data would suggest that the marking exercise did not appear to make those who attended any better able or more accurate to self-assess their future work compared with those who did not attend on closer analysis this appears to be only part of the story. Rather than making them more accurate it may be that it has opened up their horizons to what is possible and what really good work might look like, and this has had the effect of making them underestimate the quality of their own work as a result. It may also be possible that this effect is more marked on male students.

CONCLUSIONS

The continued emphasis on explicit articulation of assessment criteria and standards is not sufficient to develop a shared understanding between staff and students.

Socialisation transfer processes are necessary for tacit knowledge transfer to occur.

The traditional methods of socialisation depend on observation, imitation and feedback and discussion as the basis of personal relationship. However, this loses its effectiveness in the context of rapid expansion of student numbers and cuts in the unit of resource. It appears, however, that through a relatively simple intervention strategy incorporating a combination of explicit articulation and socialisation processes a

considerable amount may be achieved in developing shared understanding and, consequently, in improving student performance - and that this improvement may last over time and be transferable, albeit possibly only in relatively similar contexts.

BIBLIOGRAPHY

Baumard, P. (1999) *Tacit Knowledge in Organizations*, Sage Publications

Cohen, D. (1969) *Statistical power analysis for the behavioural sciences*, New York, Academic Press

Eraut, M. (1994) *Developing Professional Knowledge and Competence*, The Falmer Press, London.

Forbes, D.A. and Spence, J. (1991) An experiment in assessment for a large class, in: Smith, R. (ed) *Innovations in Engineering Education*, London, Ellis Horwood.

Gibbs, G. (1991) Reflection and self-assessment for new lecturers, in Brown, S. & Dove, P. (eds) *Self and Peer Assessment (SCED Paper 63)*, Birmingham, Standing Conference on Educational Development.

Greatorex, J. (1994) The nature and future of academic levels; the implications for the accreditation of prior learning, presented at the *Nurse Education Tomorrow Conference*, St. Mary's College, University of Durham.

Higher Education Quality Council (HEQC), 1996, *Graduate Standards Programme Draft Report and Recommendations* HEQC, London.

Hughes, I.E. (1995) Peer Assessment, *Capability*, **1**, 3, 39-43.

Laming, D. (1990) The reliability of a certain university examination compared with the precision of absolute judgements, *Quarterly Journal of Experimental Psychology Section A - Human Experimental Psychology*, **42**, 2, 239-254.

Lucas, L. and Webster, F. (1998) Maintaining standards in higher education? A case study, in: Jary, D. and Parker, M. (Eds) *The New Higher Education; issues and directions for the post-Dearing university* Staffordshire University Press, Stoke-on-Trent, 105-113.

Moon, J. (1995) *The Description of Levels in Higher Education: a discussion paper to support development of level descriptors for higher education in Wales* HECATS - Learning Outcomes Project, Wales Access Unit, Cardiff

Nelson, C.E. (1994) Critical thinking and collaborative learning, in: Bosworth, K. and Hamilton, S. (eds) *Collaborative learning and college teaching*, San Francisco, Jossey Bass.

Newstead, S. E. and Dennis, I. (1994) Examiners examined: the reliability of exam marking in psychology, *The Psychologist: Bulletin of the British Psychological Society* **7**, 216-219.

Nonaka, I. (1991) The Knowledge-Creating Company, *The Harvard Business Review*, November-December, 96-104.

O'Donovan B, Price M, Rust C (2001) 'The Student Experience of Criterion-Referenced Assessment through the Use of a Common Criteria Assessment Grid' *Innovations in Learning and Teaching International*, **38**, 1, 74-85

Otter, S. (1992) *Learning Outcomes and Higher Education* Unit for Development of Adult Continuing Education, Leicester

Price, M., O'Donovan, B. and Rust, C. (2001) Strategies to develop students' understanding of assessment criteria and processes, in: Rust, C. (ed) *Improving Student Learning: 8 Improving Student Learning Strategically*, Oxford Centre for Staff and Learning Development, Oxford.

Price, M. and Rust, C. (1999), The experience of introducing a common criteria assessment grid across an academic department, *Quality in Higher Education*, **5**, 2, 133-144.

Sambell, K, and McDowell, L. (1998), The construction of the hidden curriculum: messages and meanings in the assessment of student learning, *Assessment and Evaluation in Higher Education*, **23**, 4, 391-402

Thomas, K. (1990) *Gender and Subject in Higher Education*, Buckingham, SRHE/OUP.

Webster, F., Pepper, D., and Jenkins, A (2000) Assessing the undergraduate dissertation, *Assessment and Evaluation in Higher Education*, **25**, 1, 72-80

ACKNOWLEDGEMENT

We would like to acknowledge our debt to Roger Lindsay for his invaluable help and advice on the statistical analysis of the quantitative data.