

Student assessment from a psychological perspective

JAMES ELANDER¹

Department of Psychology, London Metropolitan University, UK

Psychological theory and research have contributed significantly to learning and teaching but have been less frequently applied to student assessment. This article reviews some of the ways in which psychology can be applied to the behaviours involved in assessment, from the formulation of assessment criteria to the reduction of biases and errors in marking. It is argued that one of the most important ways that psychology can contribute to assessment is by providing theory and methods to make implicit influences more explicit. The point is illustrated with examples of how personal construct theory has been used to develop grade descriptors and how judgement analysis has been used to investigate markers' decisions about students' work.

INTRODUCTION

Teaching, learning and assessment are all human behaviours and part of the natural subject matter of psychology. For learning and teaching, this is widely recognised (e.g. Zinkiewicz, Hammond and Trapp, 2003) and some of the most important theories of student learning draw heavily on more general psychological theories. Research on student assessment has been less closely linked to psychological theory, however, possibly because assessment involves such a wide range of behaviours that it is probably beyond the scope of any single area of psychological theory. Assessment behaviours include setting assignments, formulating assessment criteria and grade descriptors, marking students' work, providing feedback to students and moderating or agreeing marks.

Assessment takes place largely out of sight of students and without student participation, but has significant effects on student learning (Rust, 2002). For example, students may attempt to identify the 'hidden curriculum' that will be the focus of assessment (Sambell and McDowell, 1998), or prepare their work in ways that they believe will impress markers despite not being part of the assessment criteria (Norton, Dickins and McLaughlin Cook, 1996). Assessment criteria can also be used in teaching to improve student learning and achievement (Elander, 2003; Pain and Mowl, 1996; Rust, Price and O'Donovan, 2003).

Some aspects of assessment are individual activities undertaken by staff working alone, for example, when marking. Others are collective activities involving small groups of staff (moderating and agreeing marks, for example) or larger 'communities' of academics (developing assessment criteria, for example). The potential influences on those activities fall into two broad categories. Explicit influences include factors that can be clearly identified, such as formal learning outcomes and assessment criteria, or the level of agreement between markers. Implicit influences include factors that are less transparent, such as the accumulated experience and shared practice that are

not easily codified in formal language, and the tacit knowledge that informs the judgements that markers make when awarding grades. Table 1 gives examples of explicit and implicit influences on collective and individual aspects of assessment. The distinctions between these categories are not always completely clear; an individual marker working alone, for example, may be affected by shared practice in their department or institution, and many influences on assessment will have both implicit and explicit aspects. The framework is intended to serve a heuristic function in the application of psychological theory and research to different aspects of student assessment.

Table 1

Explicit and implicit influences on collective and individual aspects of student assessment

	Collective	Individual
Explicit	Agreement about assessment criteria Specification of grade descriptors	Reliability of marking Effects of training and expertise
Implicit	The 'hidden curriculum' Shared practice in marking	Tacit knowledge about standards Application of assessment criteria Judgements about grades

One approach to assessment that is consistent with recent policy and good practice is to make all the influences as explicit as possible. Precept 2 of the QAA code of practice for assessment of students is that "The principles, procedures and processes of all assessment should be explicit, valid, and reliable" (QAA, 2000, p. 6). The rights of students to information about assessment are also increasingly recognised. "Wherever possible,

¹ Correspondence concerning this article should be addressed to the author, Department of Psychology, London Metropolitan University, Calcutta House, Old Castle Street, London E1 7NT, tel +44 (0) 207 320 1073, fax +44 (0) 207 320 1236, email j.elander@londonmet.ac.uk.

students have a right to know how their essays will be marked and, generally speaking, students have a right to know afterwards the basis on which their marks were awarded" (Miller, Imrie and Cox, 1998, p. 113). There is probably a limit to the extent to which all of the influences on assessment can be made explicit, however, and it has been argued that some aspects of assessment can only be understood through shared experience (e.g. Rust *et al.*, 2003). Explicit and implicit aspects of assessment require different research strategies and one of the issues raised by the application of psychology to student assessment is the extent to which it is possible and desirable to make the process fully explicit.

This article reviews the types of behaviour that are involved in assessment and considers areas of psychological theory and research that can be applied to understand the assessment process. Two areas of recent research are described as examples of how psychological theory can be employed to make implicit influences on assessment more explicit. One is the application of personal construct theory to the formulation of grade descriptors. The second is the application of judgement analysis to markers' decisions about grades.

COLLECTIVE ASPECTS OF ASSESSMENT

Social psychology should be able to provide methods and theories to help understand how individuals work together in assessment. There are actually very few examples of research where social psychological theory has been applied directly to student assessment. This is despite the very wide range of social psychology factors that potentially influence student assessment, including the establishment and negotiation of group norms, social identities, and groupthink. Perhaps the first issue to consider is the extent to which collective activities are in fact undertaken in a collective way. As the assessment workload increases with rising student numbers (Gibbs and Lucas, 1997), there is greater motivation and potential for individuals to reduce their contribution or commitment to collective aspects of assessment. The Higher Education Quality Council (HEQC) recommended:

Methods need to be developed which ensure that the whole group of staff involved in the assessment of a programme has opportunities to discuss and compare student work and its relation to desired outcomes (HEQC, 1997, pp. 21-22).

Responses to recommendations like that could potentially draw on social psychology research on 'social loafing,' the tendency for individuals to reduce their motivation and effort when working collectively in a group (Karau and Williams, 1993).

Many of the attempts to make the assessment process more transparent and explicit involve group decision making about, for example, the assessment criteria that are to be adopted. Several departments have invested considerable time and effort in the specification of very detailed criteria that are intended to be as explicit as possible (e.g. Elander, 2002; Price and Rust, 1999).

Those efforts may be affected by the limitations of group decision making which have been the subject of a great deal of social psychology research. Group decisions about student performance should also be as transparent as possible. One study showed how group judgements about students' grades were affected by the 'common knowledge effect' whereby group discussions tend to focus on information already known to the group rather than previously unshared information. Small groups made individual and group judgements about students' grades after reading short descriptions of the students. The group judgements were influenced much more by information available to all the members of the group than by information given to just one member of the group. The effect of information distribution on group judgements appeared to be mediated by individual judgements prior to the group discussion, and group judgements were no more accurate than individual judgements (Gigone and Hastie, 1993).

It is sometimes argued that attempts to make as much as possible as explicit as possible in student assessment are misguided and devalue those aspects of assessment that can only be understood by active participation in the process. The trend to greater explicitness is sometimes criticised as reification:

If reification prevails – if everything is reified, but with little opportunity for shared experience and interactive negotiation – then there may not be enough overlap in participation to recover a co-ordinated, relevant, or generative meaning. This helps to explain why putting everything in writing does not seem to solve all our problems (Wenger, 1998, p. 65).

An alternative to 'reifying' assessment is to view it in terms of shared understandings developed through experience. Shared experience is important in the application of marking schemes because

...the precise definition of what is required does not necessarily become clear until this has been discussed at co-ordination meetings e.g. what exactly constitutes explanation rather than description? (Greatorex, Baird and Bell, 2002, p. 8).

The application of marking schemes has been described as

...a social construct negotiated by the members of the community and an individual cognitive construct (Baird, Greatorex and Bell, 2002, p. 7).

'Communities of practice' are

...tight networks or teams in which people can learn from one another yet maintain a shared ownership of the social practice (Greatorex et al., 2002, p. 4).

The concept was applied in Hall and Harding's (2002) research on the implementation of level descriptors in teacher assessment. That study distinguished between schools that exhibited many features of a 'community of assessment practice' and those where individual teachers tended to work largely in isolation from their colleagues.

Research has been conducted on the effects of factors that would be expected to help develop a community of practice among examiners, including the distribution of marking schemes, co-ordination meetings, and use of exemplar scripts. One study compared GCSE history examiners who had been randomly assigned either to a hierarchically organised meeting, a more consensual meeting, or a control group that did not meet. There were few differences between groups either in the examiners' ratings of the usefulness of the procedure (Greator *et al.*, 2002), or the actual marks they awarded (Baird, Greator and Bell, 2002). It was argued, however, that the markers involved in the study were already a sufficiently tight and experienced community of practice for the type of meeting not to make a difference (Baird *et al.*, 2002).

INDIVIDUAL ASPECTS OF ASSESSMENT

Individual behaviour is the natural unit of analysis for most psychological research, so there is much greater scope for the application of psychology to individual aspects of assessment. The activity at the heart of student assessment is the marking itself, when members of staff make judgements about the quality of each student assignment and award a grade. Most of the research on marking has focused on explicit factors such as the consistency or reliability of marking, usually in applications of psychometric and measurement theory. A more recent development is the application of psychological theory to advance understanding of implicit factors, including the tacit knowledge that markers employ when they make judgements about students' work.

Explicit individual factors

Because the task of a marker involves assigning points on a scale for students' work, measurement theory has been employed to estimate the precision of 'measurement' represented by the marking and identify sources of error or bias. Measurement theory deals mainly with two aspects of measurement; reliability and validity. Reliability refers to the stability or consistency of marking. It has been the focus of most research on marking because it can be estimated from the marks themselves when two or more marks are available for each item of work, as is the case for work that is double marked.

Validity refers to the extent to which the marks awarded reflect what the assessment was intended to measure. This is a much more difficult question to deal with than reliability, because it cannot be estimated solely from the marks awarded for each piece of work. Factors that contribute to an absence of validity can be identified if it is shown that marks were influenced by something the assessment was not intended to measure. It is much more difficult, however, to make a direct comparison between marks awarded and what should have been measured, because there is no 'gold standard' for student assessment with which the grades awarded by markers can be compared.

Estimates of the reliability of marking based on the correlation between the marks awarded to the same

work by two markers go back to the work of Hartog and Rhodes (1935) and have been continued in modern times by, for example, Laming (1990), Newstead and Dennis (1994), Dracup (1997) and Caryl (1999). The findings show that agreement between pairs of markers is highly variable and is often much lower than is considered desirable. Reliability has improved historically, possibly because of improvements in assessment practice. Reliability is much higher for degree class than for modules or units of assessment, because assessment for degree class is averaged across many units of assessment. However, the reliability of marking for individual items of work still raises questions when it shows that a substantial proportions of the variance in marks is measurement error.

Baird *et al.* (2002) suggested a number of factors that affect reliability of marking. Factors associated with the markers themselves included their expertise, training and experience. Those associated with the marking system included how closely the team of markers worked together and the types of discussion they held, whether they used exemplar work and received feedback on their marking, whether there was double marking, and their sense of ownership of the marking. Factors associated with the marking task were contrasts between the work being marked and the work marked immediately beforehand, and changes in marking practice from the beginning through the middle to the end of a batch of marking.

There is inconclusive evidence about the effects on reliability of factors associated with markers and the marking system. Baird *et al.* (2002) reported research by Weigle (1998) and Lunz, Wright and Linacre (1990) showing that training can improve the intra-rater reliability or the consistency of each individual examiner's marking. However, in one comparison between 'novice,' 'competent' and 'expert' markers, the competent and expert markers had more changes made to the grades they awarded after double marking and moderation than did the novice markers (Ecclestone, 2001).

In two controlled comparisons, neither providing markers with exemplar scripts along with marks given by the principal examiner, nor taking part in different types of examiners' co-ordination meetings had any substantial effects on the reliability of marking (Baird *et al.*, 2002). Making the marking scheme more specific may improve reliability (Lenney, Mitchel, and Browning, 1983). Price and Rust (1999) reported that, with some exceptions, the introduction of detailed assessment criteria led to improvements in marking consistency and enabled easier moderation.

Contrast effects refer to the ways in which the marks awarded for a piece of work may be affected by the quality of the work seen by the marker immediately beforehand. Spear (1997), found that good work was assessed more favourably when it followed poorer work than when it preceded it and poor work was assessed more severely when it followed better work. As Laming (1990) argued, this is consistent with the relative nature

of sensory judgements observed in laboratory experiments in psychophysical research, and may reduce the reliability of marking.

In magnitude estimation and absolute identification experiments each stimulus is compared with its predecessor because that is the only available point of reference (Laming, 1984). In the absence of a marking scheme the same will tend to happen in an examination (Laming, 1990, p. 251).

Marking research has also focused on potential biases, especially those related to student gender. Blind marking should eliminate the scope for such biases, but the issue arises for work that cannot be easily anonymised, such as student projects, where the supervisor is often one of the markers. One study of marks awarded by supervisors and second markers for student projects found that second markers gave more extreme marks to male students, so males were favoured at the top end of the scale and females at the bottom (Bradley, 1984). This was interpreted as a gender bias among the second markers. It was argued that supervisors would be less likely to be biased because they were more familiar with the students' true ability. This was consistent with the incorrect stereotype of mediocre female performance (e.g. Nicholson, 1984, p. 76). However, the same type of study conducted by Newstead and Dennis (1990) failed to find evidence of a gender bias.

Other data suggested that supervisors could be more rather than less likely to be biased in their marking of student projects. Dennis, Newstead and Wright (1996) used structural equation modelling to analyse the marks awarded to student projects. They found that approximately 30 per cent of the variance in the marks arose from factors that influenced the supervisor but not the second marker, the most likely factor being the supervisor's personal knowledge of the student. A more descriptive study of project marking also indicated that this could be affected by personal knowledge of the student. Staff reported that when marking projects they had supervised, they 'compensated' for the assessment criteria in order to take into account their perception of students' application, conscientiousness, personal pressures, personal progress and contributions during tutorials (Ecclestone, 2001).

Implicit individual factors

The most important implicit factor in marking is probably the tacit knowledge that markers and examiners bring to the judgements they make. Polanyi (1967) defined tacit knowledge as 'that which we know but cannot tell,' but it is a complex concept which has been used with different meanings in a wide range of settings (Eraut, 2000). Tacit knowledge generally refers to knowledge or expertise that is acquired through personal experience, is difficult to formalise, is not readily available to consciousness, and influences behaviour in ways that are not mediated by explicit knowledge. Eraut (2000) employed Dreyfus and Dreyfus' (1986) model of skill acquisition to argue that three types of tacit knowledge: tacit understanding, tacit procedures, and tacit rules, play important roles in the development of

professional expertise. Ecclestone (2001) argued that as individuals progress from novices to experts, they become more intuitive and less deliberative, less able to articulate the knowledge on which their decisions are based and more reluctant to deliberate because deliberation is associated with novice status.

Tacit knowledge is the rationale for the 'connoisseur' model of student assessment. This is sometimes likened to skills such as wine tasting or tea blending and is illustrated by statements such as 'I cannot describe it, but I know a good piece of work when I see it' (e.g. Rust *et al.*, 2003). The connoisseur model sometimes appears to preclude a systematic understanding of the processes involved in assessment. One approach to this problem has been to examine the shared experiences that are thought to underlie tacit knowledge in assessment, exemplified by research on 'communities of practice' in assessment (see earlier discussion of collective aspects of assessment). One limitation of communities of practice as the foundation for standards in student assessment, however, is the increasing trend towards fragmentation of academic communities in higher education, which has resulted from the need for greater explicitness in assessment.

Increasingly, assessment across different cultures and subjects involves teachers from other departments and institutions, as well as employers or professional bodies. Programmes are often part-time, modular, incorporate distance learning and assessment and increasingly use technology. Such trends challenge how far tacit notions of standards, shared in a familiar academic community which tended to take judgements on professional trust, continue to be reliable (Ecclestone, 2001).

One way that psychology should be able to contribute to understanding student assessment is by helping to make tacit knowledge more explicit. The concept of tacit knowledge is consistent with more general psychological theories of learning and memory. One explanation of tacit knowledge, for example, was in terms of Tulving's (1972) theory of episodic and semantic memory Horvath *et al.*, (1996). Also, psychological research has had considerable success in making tacit knowledge more explicit in other areas of professional life. Two areas of psychology that have been applied in this way to assessment are personal construct theory and theories of judgement and decision making.

Personal construct theory (Kelly, 1955; Bannister and Fransella, 1971) focuses on individuals' active efforts to construe or understand the world and provides a systematic theoretical framework for understanding tacit knowledge. The theory's emphasis on social processes also enables it to be related to aspects of communities of practice. For example, one of the 'corollaries' of the theory is

...to the extent that one person construes the construction processes of another, he may play a role in the social process involving another person (Bannister and Fransella, 1971, p. 30).

The research method associated with the theory is the repertory grid, which is used to allow an individual's personal understandings or constructs to emerge through the process of making comparisons. Baird (1999) argued that examiners make two types of comparison when marking; to decide whether scripts are similar to archive scripts at the same grade; and to decide whether scripts fit a personal construct of scripts at particular grades. The repertory grid technique has been adapted to investigate student assessment by eliciting information from examiners about their tacit understandings of the important differences between work at different grades. This information is then incorporated in explicit grade descriptors for the guidance of other markers and examiners.

Greatorex used repertory grids to develop grade descriptors for A-level accounting examinations (Greatorex, 2001; 2002; Greatorex, Johnson and Frame, 2001). The research involved presenting examiners with triads of examination answers and asking them in an interview to describe the ways in which the two that were awarded the same grade were similar to and different from a script with a lower grade. The approach employed a 'discriminator model' that focused on unique features of work at different grades, rather than progressive characteristics. This was based on Pollitt and Murray's (1996) previous work using Thurstone Pairs, which had shown, for example, that

...lower level performers were said to exhibit 'good grammar' but higher level performers were said to exhibit 'creativity and imagination' (Greatorex et al., 2001, p. 169).

This method was used to develop grade descriptors that provided a formal representation of the senior examiners' tacit knowledge about the qualities that distinguish performance at a particular grade (Greatorex, 2002).

Decision theory has been successfully applied to understand implicit influences on expert medical, commercial and legal judgements. It involves constructing statistical models of the associations between the judgements made by experts and a range of different 'cues' or items of specific information that influence those judgements (Einhorn, 2000). The judgement 'policy' of the expert can then be 'captured' in the form of a statistical model in which the relevant cues are weighted and combined. This provides a more explicit account of the importance attached by the expert to different aspects or cues and how they combine information about those cues. The statistical model can then be applied repeatedly and consistently to other cases involving similar judgements, eliminating errors and biases. The result of this is often that a mechanical process leads to better decisions than those based on an expert's intuitive or holistic judgement. This challenges frequent assumptions about the necessarily tacit nature of the relevant expertise. In one classic example, information from pathologists' examinations of biopsy slides were used to predict survival times for patients with Hodgkin's disease. The pathologists' overall ratings of disease severity were not related to survival times, but were

related to statistical combinations of several characteristics of each slide (Einhorn, 1972).

The situation in student assessment is analogous in that markers must make an overall judgement about the quality of students' work in which a number of specific cues – the assessment criteria – should be represented. The difference between student assessment and clinical judgement, however, is that for student assessment there is no equivalent of survival time against which to compare the markers' judgements. There is no 'gold standard' for student performance, however, the method can still be used to 'capture' individual marking policies.

Elander and Hardman (2002) conducted a study in which markers made separate ratings of seven assessment criteria for examination answers, as well as awarding an overall mark for each answer. This allowed each marker's implicit 'assessment policy' to be made more explicit, by examining the ways in which the ratings for specific criteria were related to the overall mark awarded. The most interesting aspects of the results concerned differences between the first markers who had taught the material being examined and set the question, and second markers who were members of staff with more general expertise in the area of the examination. The overall marks awarded by first markers were related to markers' ratings on six of the seven assessment criteria, whereas those awarded by second markers were related to only three. They were particularly related to the criterion of 'covering the area'. It appeared that first markers, having taught the material and set the questions, were more able than second markers to incorporate criteria such as 'evaluates the material' and 'develops arguments' in the marks they awarded. This is consistent with the observation that

...experts in a particular topic were more likely to see analysis in students' work compared to novice assessors in the topic who could not easily detect subtleties of argument or, conversely, a simplistic treatment of issues (Ecclestone, 2001, p. 309).

CONCLUSIONS

The range of behaviours that are involved in assessment are unlikely to be integrated by any single area of psychology, but almost all are potentially amenable to the application of at least one area of research or theory. The most promising aspects of assessment from this point of view are those that are traditionally regarded as implicit. Implicit factors are by definition those requiring further explanation, so it is not a surprising conclusion that these should be the main focus for further research. What is perhaps surprising is the relative lack of attention that psychologists have paid to this aspect of assessment, because psychology has had significant success in making more explicit the influences on other types of complex judgement. In medicine, for example, applications of decision theory and judgement analysis have made substantial contributions to improving the diagnosis and treatment of illness.

It might be argued that because tacit knowledge is based in experience, it can only be understood through shared activity, practice and experience. This is almost to suggest that certain aspects of professional practice in student assessment are inherently mysterious or unknowable and that attempts to understand them in explicit terms could not succeed. Against this is the fact that psychology has achieved exactly that in a number of other areas. Indeed, it is in the nature of psychology that it should be capable of advancing understanding of aspects of human behaviour about which the protagonists have limited introspective insight and where subjective accounts may be misleading or self-serving. The record of applied psychological research in the workplace, in medicine, and indeed in almost every other aspect of education, strengthens the claim that psychology should be able to contribute to understanding and improving student assessment. The claim is further strengthened by the results of the work that has been conducted so far in this area, which have led to more authentic grade descriptors and insights into marking judgements.

It might alternatively be argued that although psychology could help to articulate what is presently implicit, it could never provide a complete account of what is involved in assessment. This is because such an account will always include elements of shared experience and tacit knowledge that will resist explication. This is very probably true, at least while student assessment continues to be conducted by communities of staff working closely together, so that psychology can only provide a *contribution* to improving understanding and improving practice. Student assessment has many aspects and both practitioners and researchers must find a balance between different approaches. Where aspects of assessment do become more explicable through careful research, however, we should recognise that the balance between implicit and explicit factors can change, at least temporarily. There is no advantage in attempting to preserve the tacit status of practices that have become better understood than before. Knowing more in an explicit way about certain aspects of assessment should allow us to focus our use of implicit approaches more effectively, for example by sharing active subjective experience of those aspects of assessment that remain poorly understood in explicit terms.

Nor is it truly self-serving for academics to seek to preserve the mystique of implicit knowledge about student assessment. Research like that of Greatorex on the development of grade descriptors supports professional practice by providing descriptors that captured more clearly the understandings of expert practitioners about the meanings of grades. Research like that of Elander and Hardman can help to guide practice in moderating and agreeing marks by highlighting ways in which markers may have different perspectives in their readings of examination answers. Any of the research findings described in this article could provide the stimulus for reflective debate among practitioners. This debate can feed into the shared and tacit experience of assessment, because those aspects

need not be insulated from those that can be made more explicit. The distinctions drawn in this article between different aspects of assessment serve the heuristic purpose of organising material more than they represent natural divisions between distinct phenomena. It is also worth knowing more, and more explicitly, about assessment, because of the influence of assessment on student learning. The more that can be understood about assessment in explicit terms, the more can be explained to students in ways that improve learning and achievement. Despite recent initiatives to include students in the community of assessment practice to which markers and examiners belong (e.g. Rust *et al.*, 2003), there are limits to how far students can learn tacitly about assessment through experiences shared with staff.

A recent trend in student assessment has been the production of more detailed assessment criteria and this could provide a focus for further research. Some of those provide specific criteria for aspects of essays and examination answers such as addressing the question, using evidence, developing arguments, structuring material and so on (Elander, 2002; Price and Rust, 1999). One aim was to improve students' understanding and achievement, for which there was some evidence of success (Rust *et al.*, 2003) and another was to support more valid and reliable marking, about which there is presently little evidence.

The most difficult issue in student assessment, however, is the validity of marking. There is no gold standard or external criterion for the overall quality of an essay or examination answer, so the mark reflects a judgement in which various aspects of the work are weighed in the mind of the marker. It is possible that aspect-specific assessment criteria may open the door to studies of the validity of marking. Separate ratings by markers of the extent to which an essay answered the question, developed arguments, used evidence and so on, could be compared with more objective information from the essay itself. A content analysis of essays could potentially identify the parts of the essay that contributed to meeting the various aspect-specific criteria. Research of this type would answer questions about markers' ability to identify and evaluate specific aspects of students' work, the extent to which different specific criteria are in fact independent of one another, and how specific aspects of students' work influence the judgements made by markers. Such information could lead to the development of more useful assessment criteria, together with concrete examples of how the criteria were demonstrated in students' work. Those examples could then be used to explain to students what is meant by certain criteria that students often find difficult to understand, for example 'critical evaluation'.

More ambitiously, research of this type could form the basis for alternative assessment methods that employ specific rather than global assessments. One study seemed to show that second markers were good at rating specific aspects of students' examination answers, but less good at incorporating those ratings in an overall mark for the answer (Elander and Hardman,

2002). Research on aspect-specific assessment criteria could lead to assessment methods using an algorithm to combine markers' ratings of specific aspects of the work rather than relying on the marker to make an overall judgement. This is similar to the way in which expert judgement has been modelled in medicine and business.

In a recent essay on tacit knowledge in professional life, Eraut commented:

There is a danger that the continuing discovery of the importance of tacit knowledge will lead some people to argue on ideological grounds that it should replace evidence-based practice. My own view is the opposite, that we should seek to expand evidence-based practice but not suffer from any delusions about how far it will take us nor lose awareness of just how much interpretation of guidelines may be needed when making decisions about individual cases (Eraut, 2000, p. 125)

Eraut did not specifically consider assessment, but those conclusions serve equally well as recommendations for the application of psychology to student assessment.

REFERENCES

- Baird, J. (1999, December). *Are examination standards all in the head? Experiments with examiners' judgements of standards in A level examinations*. Paper presented at the British Psychological Society Conference, Institute of Education, London
- Baird, J., Greatorex, J. and Bell, J. F. (2002, September). *What makes marking reliable? Experiments with UK examinations*. Paper presented at the International Association for Educational Assessment Conference, Hong Kong.
- Bannister, D. and Fransella, F. (1971). *Inquiring Man: the Theory of Personal Constructs*. Harmondsworth: Penguin.
- Bradley, C. (1984). Sex bias in the evaluation of students. *British Journal of Social Psychology*, 23, 147-163.
- Caryl, P. G. (1999). Psychology examiners re-examined: a 5-year perspective. *Studies in Higher Education*, 24, 61-74.
- Dennis, I., Newstead, S. E. and Wright, D. E. (1996). A new approach to exploring biases in educational assessment. *British Journal of Psychology*, 87, 515-534.
- Dracup, C. (1997). The reliability of marking on a psychology degree. *British Journal of Psychology*, 88, 691-708.
- Dreyfus, H. L. and Dreyfus, S. E. (1986). *Mind over machine: the power of human intuition and expertise in the area of the computer*. Oxford: Basil Blackwell.
- Ecclestone, K. (2001). "I know a 2.1 when I see it": understanding degree standards in programmes franchised to colleges. *Journal of Further and Higher Education*, 25, 301-313.
- Einhorn, H. J. (1972). Expert judgement and mechanical combination. *Organizational Behaviour and Human Performance*, 7, 86-106.
- Einhorn, H. J. (2000). Expert judgement: some necessary conditions and an example. In T. Connolly, H. R. Arkes and K. R. Hammond (Eds.), *Judgement and Decision Making: an Interdisciplinary Reader* (2nd edition) 324-335. Cambridge: Cambridge University Press.
- Elander, J. (2002). Developing aspect-specific assessment criteria for examinations and coursework essays in psychology. *Psychology Teaching Review*, 10, 31-51.
- Elander, J. (2003). A discipline-based undergraduate skills module. *Psychology Learning and Teaching*, 3, 48-55.
- Elander, J. and Hardman, D. (2002). An application of judgement analysis to examination marking in psychology. *British Journal of Psychology*, 93, 303-328.
- Eraut, M. (2000). Non-formal learning and tacit knowledge in professional work. *British Journal of Educational Psychology*, 70, 113-136.
- Gibbs, G. and Lucas, L. (1997). Coursework assessment, class size and student performance: 1984-94. *Journal of Further and Higher Education*, 21, 183-192.
- Gigone, D. and Hastie, R. (1993). The common knowledge effect: information sharing and group judgement. *Journal of Personality and Social Psychology*, 65, 959-974.
- Greatorex, J. (2001). Making the grade – how question choice and type affect the development of grade descriptors. *Educational Studies*, 27, 452-464.
- Greatorex, J. (2002). Making accounting examiners' tacit knowledge more explicit: developing grade descriptors for an accounting A-level. *Research Papers in Education*, 17, 211-226.
- Greatorex, J. Baird, J. and Bell, J. F. (2002, August). *'Tools for the trade': what makes GCSE marking reliable?* Paper presented at the Learning Communities and Assessment Cultures: connecting research with practice, University of Northumbria, Newcastle upon Tyne. Retrieved 26 February 2004 from <http://www.uclcs-red.cam.ac.uk/conferencepapers.htm>.
- Greatorex, J., Johnson, C. and Frame, K. (2001). Making the grade – developing grade descriptors for accounting using a discriminator model of performance. *Westminster Studies in Education*, 24, 167-181.
- Hall, K. and Harding, A. (2002). Level descriptors and teacher assessment in England: towards a community of practice. *Educational Research*, 44, 1-16.
- Hartog, P. and Rhodes, E. C. (1935). *An Examination of Examinations*. London: MacMillan.
- Higher Education Quality Council (HEQC) (1997). *Assessment in Higher Education and the Role of 'Graduatness'* London: Higher Education Quality Council.
- Horvath, J. A., Sternberg, R. A., Forsythe, E. B., Bullis, R. C., Williams, W. M. and Sweeney, P. J. (1996). *Implicit theories of leadership practice*. Paper presented at the Annual Meeting of the American Educational Research Association (AERA), New York.
- Karau, S. J. and Williams, K. D. (1993). Social loafing: a meta-analytic review and theoretical integration. *Journal of Personality and Social Psychology*, 65, 681-706.
- Kelly, G. A. (1955). *The Psychology of Personal Constructs*. New York: Norton.
- Laming, D. (1984). The relativity of 'absolute' judgements. *British Journal of Mathematical and Statistical Psychology*, 37, 152-183.

- Laming, D. (1990). The reliability of a certain university examination compared with the precision of absolute judgements. *Quarterly Journal of Experimental Psychology*, 42A, 239-254.
- Lenney, E., Mitchel, L. and Browning, C. (1983). The effect of clear evaluation criteria on sex bias in judgements of performance. *Psychology of Women Quarterly*, 7, 313-327.
- Lunz, M. E., Wright, B. D. and Linacre, J. M. (1990). Measuring the impact of judge severity on examination scores. *Applied Measurement in Education*, 3, 331-345.
- Miller, A. H., Imrie, B. W. and Cox, K. (1998). *Student assessment in higher education: a handbook for assessing performance*. London: Cogan Page.
- Newstead, S. E. and Dennis, I. (1990). Blind marking and sex bias in student assessment. *Assessment and Evaluation in Higher Education*, 15, 132-139.
- Newstead, S. E. and Dennis, I. (1994). Examiners examined: the reliability of exam marking in psychology. *The Psychologist*, 7, 216-219.
- Nicholson, J. (1984). *Men and women: how different are they?* Oxford: Oxford University Press.
- Norton, L. S., Dickins, T. E. and McLaughlin Cook, A. N. (1996). Rules of the Game in essay writing. *Psychology Teaching Review*, 5, 1-14.
- Pain, R. and Mowl, G. (1996). Improving geography essay writing using innovative assessment. *Journal of Geography in Higher Education*, 20, 19-31.
- Polanyi, M. (1967). *The Tacit Dimension*. Garden City, NY: Doubleday.
- Pollitt, A. and Murray, N. L. (1996). What raters really pay attention to. In M. Milanovic and N. Saville (Eds.), *Studies in language testing 3: performance testing, cognition and assessment: selected papers from the 15th language testing research colloquium*. Cambridge and Arnhem: Cambridge University Press.
- Price, M. and Rust, C. (1999). The experience of introducing a common criteria assessment grid across an academic department. *Quality in Higher Education*, 5, 133-144.
- Quality Assurance Agency (QAA) for Higher Education (2000). *Code of practice for the assurance of academic quality and standards in higher education. Section 6: assessment of students*. Gloucester: QAAHE. Retrieved 26 February 2004 from http://www.qaa.ac.uk/public/cop/copaosfinal/COP_AOS.pdf.
- Rust, C. (2002). The impact of assessment on student learning: how can research literature practically help to inform the development of departmental assessment strategies and learner-centred assessment practices. *Active Learning in Higher Education*, 3, 145-158.
- 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, 147-164.
- 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, 391-402.
- Spear, M. (1997). The influence of contrast effects upon teachers' marks. *Educational Research*, 39, 2, 229-233.
- Tulving, E. (1972). Episodic and semantic memory. In E. Tulving and W. Donaldson (Eds.), *Organisation of memory*. New York: Academic Press.
- Weigle, S. (1998). Using FACETS to model rater training effects. *Language Testing*, 15, 263-287.
- Wenger, E. (1998). *Communities of practice: learning, meaning and identity*. Cambridge: Cambridge University Press.
- Zinkiewicz, L., Hammond, N. and Trapp, A. (2003). *Applying Psychology Disciplinary Knowledge to Psychology Teaching and Learning*. Report and Evaluation Series No 2. York: LTSN Psychology, University of York.

Manuscript received on 30 April 2003

Revision accepted for publication on 15th December 2003