in formative assessment may decrease validity in summative assessment. Furthermore, the simple knowledge that the result is being used for one purpose (e.g. school league tables) may decrease its validity for another. But, this said, there is no reason why an assessment should not serve a number of different purposes, so long as we are clear what these are, and where our priorities lie.

Standardisation is about standards, and there is an ongoing debate over whether standards, for example in A-levels, are going up or down. To get a grip on this we need to consider what is meant by 'standards'. For example, teaching standards are not the same as the standard of achievement. It is perfectly possible for standards of teaching to go up at the same time as standards of achievement go down, and vice versa. Also, standards are not necessarily applicable across the board. A form of teaching that raises standards for one group (for example, children with special educational needs) may lower them for another.

The desire to design assessments, examinations and tests that are free from bias is as much a concern for school examining bodies as it is for recruitment professionals. Unfortunately, given the existence of extrinsic test bias, assessment that is completely free from bias is in many cases an impossibility. But we can all endeavour to keep bias to a minimum, and to do so is an important part of any equal opportunities policy, whether that of an organisation or enshrined in law within equal opportunities legislation. What is important is that its extent should be monitored and discussed, and that programmes to evaluate and reduce its extent should be incorporated in policy. This can be difficult where companies and organisations are in denial, and it will be an uphill task to ensure that the issue receives the attention it deserves. As far as A-levels are concerned, two forms of bias are apparent. First, the differences in attainment between ethnic groups, and secondly, the superior performance of girls compared with boys, in some subjects. As far as ethnic groups are concerned, the differences in quality of schooling between inner cities and the suburbs is sufficiently manifest not to need much discussion, although the causes of these differences are of course a different matter. One thing we can be sure of, however, is that attempts to deflect the issue on to universities are unlikely to lead to the changes we need. The black and Bangladeshi communities in particular deserve to have their concerns in this respect recognised and addressed.

With gender differences in achievement, it is interesting to note that several decades ago boys outperformed girls at A-level, a situation that is now reversed. Is this because girls are now cleverer than boys? Not necessarily. Two other elements will almost certainly have come into play. First is the higher standard deviation for boys compared with girls on most ability and achievement tests. This generally means that boys are over-represented at the extremes of the distribution. A shift in the cut-off closer to the population average, as effectively happens when the participation rate shifts from 10% to 50%, could very easily show that the previous superior performance of boys was an artefact. A second change in the way A-level is examined will also have contributed, this being the increased dependence of the final mark on coursework. There are complex interactions between gender and various aspects of the coursework process.

The psychometric principles are not new, and necessarily underlie much of the activities of examination boards in their efforts to improve the culture of learning, examinations and the monitoring of performance. They are also inescapable, although sometimes attempts are made to dress them up in other clothes. Perhaps this is inevitable given the increasing politicisation of our school system. Is it too much to hope that one day the curriculum and its assessment will be disestablished? The freedom given to the Bank of England to set interest rates independent of Treasury interference has set a useful precedent here. Only time will tell.

VOCATIONAL ASSESSMENT

Is passing just enough? Some issues to consider in grading competence-based assessments

Martin Johnson Research Division

Introduction

Competence-based assessment involves judgements about whether candidates are competent or not. For a variety of historical reasons, competency-based assessment has had an ambivalent relationship with grading (i.e. identifying different levels of competence), although it is accepted by some that 'grading is a reality' (Thomson, Saunders and Foyster, 2001, p.4). The question of grading in competence-based qualifications is particularly important in the light of recent national and international moves towards developing unified frameworks for linking qualifications. This article is based on Johnson (2006, in submission) which uses validity as a basis for discussing some of the issues that surround the grading of competence-based assessments. The article is structured around 10 points taken from the summary of that extended paper.

1. Defining competency

This can be problematic and might be conceptualised in terms of atomistic/holistic or tacit/instrumental factors. Competency-based assessment systems have developed in the context of these varying conceptualisations.

The assessment systems used to represent and measure competent performance are inextricably tied to the ways that 'competence' has been defined. Debates about the nature of competence have tended to be polarised around the question of whether it is a complex or superficial construct, with consequent implications for assessment methods. Wood (1991) cites literature highlighting the inherent difficulties of inferring competence from test data or observed performance. He suggests that this is partly because those constructs that might be regarded by some

as contributing to a notion of competency are often grossly under-conceptualised. This potentially leads assessment-based inferences about competence to be invalidly 'over-extended'.

More sophisticated conceptualisations tend to consider those attributes that underpin performance. Gonczi (1994) outlines a broad model of competence that prioritises the personally held skills which, in common, underpin competent performance. Gillis and Bateman (1999) also acknowledge a broader conception of competency, arguing that competency must include the application of skills across contexts (location and time), and the generic transferable skills, sometimes referred to as 'key skills', that enhance the capacity of workers to respond, learn and adapt when environmental factors change.

There are also concerns about whether competence can be satisfactorily defined and the role of assessor experience in judgements about competent performance. Some argue that attempts to over-specify detailed assessment criteria in order to attain unambiguous, reliable judgements might not have the desired outcome. Wolf (1995) observes that written specifications on their own might well leave space for ambiguous interpretation since no criterion, however precisely defined, is beyond multiple interpretations. Although it appears counterintuitive to suggest that very detailed assessment criteria may leave space for personal interpretation, when faced with a mass of criteria an assessor may well read through them and glean a sense of meaning, perhaps giving their own weight to particular points and therefore reducing the overall consistency of application.

Others argue that attempts to over-specify 'transparent' assessment criteria will also have limited success because of the particular influence of tacit knowledge in competent performance. Situated cognition theorists suggest that the development of competence involves 'knowing in practice' and becomes embodied in the identity of the practitioner (Lave and Wenger, 1991). Modelling the different stages of developing expertise, Dreyfus and Dreyfus (1986) argue that tacit, intuitive understanding is a critical difference between the performances of experts and novices.

2. Grading and motivation

There is considerable debate about the potential advantages and disadvantages of grading on motivation. Literature suggests that the reporting of performance outcomes can influence learner motivation. Social Cognitive theorists, such as Bandura (1986), hold that individuals use feedback from past experiences (successes and failures) to inform their expectations about future performance. Perhaps unsurprisingly, the quality of this information can affect perceptions of self-efficacy and influence future motivation to act.

Grading potentially gives more feedback about performance than binary reports. As a consequence, Smith (2000) suggests that grading can facilitate the motivation for students to strive for excellence since the reporting mechanism affords the opportunity for this level of performance to be recognised.

There is evidence that the effects of grading are not consistent across all learners. Williams and Bateman (2003) suggest that whilst more able learners might consider grading to be more motivational because it recognises their strengths, lower ability learners might be adversely affected. It is also important to consider the potential relationship between grading and labelling. There are concerns that learners might internalise the descriptive quality attached to grades to the extent that they infer that their performance (and ability) is a fixed, unchangeable entity.

The nature of learners who take vocational courses might be different from those who opt for general qualifications, and their motivation might differ. Group dynamic issues might also need consideration. Usually vocational learning takes place in smaller groups than is the case for general learning. This might contribute to a greater sense of group cohesion, undermining the motivation of individuals to compete against their peers.

3. The effects of grading on (mis)classification

Smith (2000) asserts that grading can improve the validity and consistency of assessments because it compels assessors to analyse students' performances with greater care than in binary reporting systems. This might be because they have to consider the evidence of a performance at a finer grain. On the other hand, this will only be possible if the inherent logic of the subject provides recognisable thresholds (Wolf, 1993).

Williams and Bateman (2003) highlight the potential relationship between the number of grading boundaries and the reliability of assessment outcomes. The opportunity for classification errors increases simply because the number of differentiated classifications increases. However, the errors might have less severe consequences. Overcoming this problem could potentially undermine the consistent reporting of outcomes since it demands greater levels of accuracy in each assessment judgement. Newton (2005) argues that the existence of measurement inaccuracy impacts on the social credibility of assessments because of public expectation that there should be relatively few misclassifications.

Finally, Wiliam (2000) emphasises the danger of aggregating marks into grades or levels since these might mask the true extent of error variance in test scores. Since the exactness of test scores can give an illusion of precision, resulting in misleading perceptions about their real accuracy, grading might be considered more favourable because it suffers less from this degree of definition.

4. Stretching assessment criteria beyond binary outcomes

Another important consideration is the interaction between domain breadth and the constructs included. Disentangling these interacting factors allows a clearer discussion regarding the potential consequences of grading. Domains can often be broad, requiring the integration of a number of identifiable skills. This raises questions about the nature of competent performance, since the term might be used (and understood) in different senses. Hyland (1994) suggests that competence might be both a holistic evaluation against a professional standard (e.g. being a competent plumber) and an atomistic evaluation of the ability to achieve a particular task (e.g. a particular driving manoeuvre). In the first context he argues that grading is appropriate because the holistic nature of the performance might include observable degrees of performance. However, in the second context grading might be inappropriate because atomistic tasks might not be scalable beyond 'achieved' and 'not yet achieved'.

5. Grading and accountability

There are concerns that grading procedures afford comparisons to be made between institutions and that these can be used for accountability purposes. In this way grading can be a potential source of pressure for

assessors and might influence their decision-making. Wikström (2005) explores some of the structural pressures beyond the immediate context of assessment tasks which can impact on the integrity of grading decisions in a criterion-referenced system. She found evidence that teachers' and tutors' grading decisions were affected by selection and accountability concerns. Her findings suggest that teachers might grade differently over time because of:

> *Both internal and external pressures for high grading, due to the grades' function as a quality indicator for schools as well as a selection instrument for students*. (p.126)

Similarly, Bonnesrønning (1999) posits a systematic relationship between teacher characteristics, such as self-confidence levels, and grading practices. For example, he states that:

> *Teachers' ability to withstand pressure [for high grading] varies with teacher characteristics*. (p.103)

This could have implications for perceptions about the robustness of teacher or tutor assessed competency reports.

6. Decisions about grading depend on the domain being assessed

Decisions about grading need to consider the context of the domain being assessed. Grading decisions should be based on the number of usefully distinct subject specific criteria which can be formulated, the inherent logic of the subject, and whether there are recognisable thresholds. Messick (1989) argues that consideration of the consequences of assessment results is central to validity, stating that:

> *Validity is an integrated evaluative judgement of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment*. (p.13)

Considering the interpretation of assessment evidence leads to a focus, amongst other things, on the quality of the information gained from an assessment. Although considering whether it is validly possible to separate binary into graded outcomes is important, arguably domains traditionally used for binary judgements grading can offer additional information. This could afford more information on which to base inferences about individual achievement and contribute to the validity of the assessment process. It assumes that inferences about competence are based on a sound understanding of the grading criteria. Transparency about how grades are determined is important. However, the meaning of different grade thresholds is less transparent than that between competent/not competent if there is a lack of understanding about the differences between grades.

7. Context and norm-referenced interpretations

Literature suggests that context and norm-referenced interpretations might undermine the validity of applying grading procedures to competency-based assessments. Context might interfere with consistency in at least two ways. First, context might interfere with an assessor's ability to position the qualities of two different performances on a common scale. Factors may exist that intrude on the process of casting consistent judgements (e.g. performances in tasks involving interactions between individuals might be interpreted differently by judges who accommodate variations in the social dynamics, such as, dealing with 'tricky' as opposed to 'helpful' customers). Secondly, context can make it more difficult to infer the basis on which assessors' decisions are being made. Assessors in different contexts might make judgements based on different foundations from each other because their understanding of competence is based on their different experiences.

Where binary reporting methods are used there is a clear, transparent link relating pass/fail distinctions to particular criteria. One of the problems for competency-based assessment is that qualification users might mistakenly assume that graded performance reporting is based on norm-referenced principles. Williams and Bateman (2003) and Peddie (1997) found that qualifications stakeholders sometimes make this mistake.

However, a number of commentators questioned whether criterion-referenced judgements are entirely devoid of norm-referenced principles. Skidmore (2003) argues that criteria could be based on an underlying normative judgement where they rely on subjective interpretation by professional judges. Similarly, Wiliam (1998), citing Angoff (1974), suggests that any criterion-referenced assessment is underpinned by a set of norm-referenced assumptions because the assessments are used in social settings, and assessment results are only relevant with a reference to a particular population. Consequently, any criterion-referenced assessment is attached to a set of norm-referenced assumptions.

8. The use of 'merit' grades

Using grading in competency-based assessments might demotivate and discourage some learners. One method of overcoming this problem is to grade outcomes once competence has been established. 'Merit' and 'excellence' grades might be used for this purpose, although Peddie (1997) suggests that these terms need to be distinguished so that they are used validly. According to Peddie, 'merit' and 'excellence' have different qualities; 'excellence' has an exclusivity, implying that some students are excellent in relation to a larger group of students who are not excellent, whilst 'merit' means very good, potentially being attained by all students. In this context, 'merit' grading can help to identify praiseworthy performances, without necessarily engaging the norm-referenced techniques that some argue undermine competency-based assessment principles.

9. Grading potentially affords the use of assessment data for selection purposes

An important use of assessment outcomes is to inform selection decisions, Wolf (1995) states a commonly held view that:

> *In a selection system, a simple pass/fail boundary provides far too little information on which to base decisions*. (p.75)

Grading can perform an important role where decisions need to be made about selection or access to limited opportunities and/or resources. Fewer grades will result in fewer fine distinctions between performance descriptions. The social consequences of this might be selectors placing a greater emphasis on other selection criteria, which might be less reliable than the examination/assessment itself. In addition, it reduces the effect of measurement error. For example, a pass might be a misclassified, incompetent applicant but an excellent result is much less likely to be one.

10. Grading can help to establish the comparative status of different qualifications

Grade creation based on the distribution of performances within the population can be one way of enabling comparisons between assessments to be made. It is also important to acknowledge that grading might encourage the use of particular frames of understanding which look to make comparisons across domains. The creation of graded performance scales might encourage the development of common assumptions about the similarity of skills and demands that are needed to achieve similar grades across different domains. The extent to which this is possible and valid is questionable although the construction of such comparisons is notionally encouraged by the grading framework.

A consequence of using grades as a tool for comparing the vocational and academic domains is the potential for 'a paradox of parity' (Griffin and Gillis, 2001). An important function of Vocational Education and Training (VET) is to encourage less academic students to remain at school. However, in order to achieve parity of esteem and intellectual demand with other 'academic' subjects, there is a perceived need to attract more academically able students into those vocational subjects. A paradox of parity could occur if this is successful since less able students might be discouraged from enrolling in VET courses which appear increasingly similar to 'academic' courses.

Conclusion

In theory, grading can be an appropriate method for dealing with ordinal competency assessment data, although there are claims that data from competency-based assessments should be regarded as being nominal. In practice, the potential benefits of grading need to be balanced against its potential disadvantages. This article suggests that questions about the desirability of grading competency-based assessments are related to issues of validity, with the question hinging on the simultaneous existence of two mutually supporting factors: 'use value' and 'validity'. It appears that the grading of competence-based assessments can only be justified where both factors exist, in other words where it has a clear value for qualification users and where its application is valid. The existence of either of these factors in isolation undermines the use of grading since it weakens the crucial link between the generation of sound assessment data and its complementary interpretation.

References

Angoff, W. H. (1974). Criterion-referencing, norm-referencing and the SAT. *College Board Review*, **92**, 2–5.

Bandura, A. (1986). *Social foundations of thought and action: A Social Cognitive theory*. Englewood Cliffs, NJ: Prentice-Hall.

Bonnesrønning, H. (1999). The variation in teachers' grading practices: Causes and consequences. *Economics of Education Review*, **18**, 1, 89–105.

Dreyfus, H. L. & Dreyfus, S. E. (1986). *Mind over machine: The power of human intuition and expertise in the age of the computer*. Oxford: Basil Blackwell.

Gillis, S. & Bateman, A. (1999). *Assessing in VET: Issues of reliability and validity*. Kensington Park, South Australia: NCVER.

Gonczi, A. (1994). Competency-based assessment in the professions in Australia. *Assessment in Education*, **1**, 1 27–44.

Griffin, P. & Gillis, S. (2001). *Competence and quality: Can we assess both?* Paper presented at the National Conference on Grading and Competency Based Assessment in VET, Melbourne, May, 2001.

Hyland, T. (1994). *Competence, education and NVQs: Dissenting perspectives*. London: Cassell Education.

Johnson, M, (2006). Grading in competence-based qualifications: Is it desirable? *The Journal of Further and Higher Education*. In submission.

Lave, J. & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. Cambridge: Cambridge University Press.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement*, 13–103. Washington DC: American Council on Education/Macmillan.

Newton, P. E. (2005). The public understanding of measurement inaccuracy. *British Journal of Educational Research*, **31**, 4, 419–442.

Peddie, R. A. (1997). Some issues in using competency based assessments in selection decisions. *Queensland Journal of Educational Research*, **13**, 3, 16–45. http://education.curtin.edu.au/iier/qjer/qjer13/peddie.html

Skidmore, P. (2003). *Beyond measure: Why educational assessment is failing the test*. London: DEMOS.

Smith, L. R. (2000). *Issues impacting on the quality of assessment in vocational education and training in Queensland*. Brisbane: Department of Employment, Training and Industrial Relations.

Thomson, P., Saunders, J. & Foyster, J. (2001). *Improving the validity of competency-based assessment*. Kensington Park, SA: NCVER.

Wikström, C. (2005). Grade stability in a criterion-referenced grading system: the Swedish example. *Assessment in Education*, **12**, 2, 125–144.

Wiliam, D. (2000). Reliability, validity, and all that jazz. *Education 3–13*, **29**, 3, 9–13.

Wiliam, D. (1998). *Construct-referenced assessment of authentic tasks: Alternatives to norms and criteria*. Paper presented at the 24th Annual Conference of the International Association for Educational Assessment, Barbados.

Williams, M. & Bateman, A. (2003). *Graded assessment in vocational education and training*. Kensington Park, South Australia: NCVER.

Wolf, A. (1993). *Assessment issues and problems in a criterion-based system*. London: Further Education Unit, University of London.

Wolf, A. (1995). *Competence-based assessment*. Buckingham: Open University Press.

Wood, R. (1991). *Assessment and testing: A survey of research*. Cambridge: University of Cambridge Local Examinations Syndicate.