Association Annual Conference, Edinburgh, September 2003. http://www.cambridgeassessment.org.uk/research/confproceedingsetc/BERA2003JBEMMS/file/

Bell, J.F., Malacova, E. & Shannon, M. (2005). The changing pattern of A level/AS uptake in England. *The Curriculum Journal*, **16**, 3, 391–400.

Clarke, P. (2005). *Curriculum 2000 and other qualifications. A survey of UK medical schools' requirements and selection policies*. Cardiff University. http://www.ucas.ac.uk/candq/curr2000/medical.pdf

Centre for Education and Industry (2003). *Widening participation in the physical sciences: An investigation into factor influencing the uptake of physics and chemistry*. University of Warwick, Coventry: Centre for Education and Industry.

Robson, W. W. (1965). *English as a university subject*. London: Cambridge University Press.

Smithers, A. & Robinson, P. (2006). *Physics in schools and universities: II. Patterns and policies*. Buckingham: Centre for Education and Employment Research. http://www.buckingham.ac.uk/education/research/ceer/pdfs/physicsprint-2.pdf

Tillyard, E.M.W. (1958). *The muse unchained: An intimate account of the revolution in English Studies at Cambridge*. London: Bowes and Bowes.

Vidal Rodeiro, C.L. (2006). Uptake of GCE A-level subjects in England 2001–2005. Statistical Report Series No. 3. http://www.cambridgeassessment.org.uk/research/statisticalreports/

Discussion piece: The psychometric principles of assessment

Professor John Rust Psychometrics Centre

Psychometrics is the science of psychological assessment, and is a foundation of assessment and measurement. Within psychometrics there are four fundamental principles whereby the quality of an assessment is judged. These are (1) reliability, (2) validity, (3) standardisation and (4) freedom from bias. Reliability is the extent to which an assessment is free from error; validity is the extent to which a test or examination assesses what it purports to assess; standardisation gives us information on how the result of an assessment is to be judged, and freedom from bias examines the extent and causes of differences between groups. These four principles inform not only test use but also the entire process of test development, from the original curriculum or job specification, via the choice and appraisal of examination questions and test items, through to the eventual evaluation of the success or otherwise of the assessment itself.

No assessment can be perfectly reliable, and this applies not only to the measurements we make in education or psychology, but to all types of measurement. Measurements range in accuracy from the exceptionally high levels now obtained for the speed of light and the time of day, through measurements of length and area used in surveying, to the lower levels attainable for measurement of blood pressure and haematological assays used in medicine, to the tests of ability, achievement and character with which we are familiar in the education and recruitment testing world. Hence, in all these cases our expectations are different. Reliability is assessed on a scale of zero to one, with a score of 0.00 indicating no reliability at all, and a score of 1.00 representing perfect reliability. Over a century of human testing has shown us that we can expect reliabilities ranging from 0.95 for a very carefully constructed and individually assessed test of ability, through 0.85 for group tests of ability; about 0.75 for personality tests; 0.5 for rating scales and down to about 0.2 or 0.3 for projective tests or tests of creativity.

There are several ways in which reliability can be assessed and most of them involve making multiple measurements. Inter-rater reliability is the extent to which examiners agree or disagree about the mark that a candidate should be given when the assessments are made independently. This is the most relevant form of reliability for many forms of school examinations, and the assessment of coursework or essays has an apparent upper limit of about 0.65. It is interesting that in spite of frequent attempts to improve on inter-relater reliability of examiners, for example by setting detailed marking criteria, it is unusual to find a value that goes much above this figure. Constraining the judgement of individual examiners can, if taken to extremes, lead to just another multiple choice test by another name.

Whatever efforts are put into improving the agreement between markers, it is not the only source of unreliability and may not even be the most important. In multiple choice examinations, for example, the inter-rater reliability is often as high as 0.99, because the only disagreement between raters is in reading the scores from the answer sheets. However, unreliability still arises for many other reasons, such as the state of the candidate (tired, ill, anxious etc.) the environment in which the test is taken, events at home or among the peer group, or the concordance between the content of the examination and the revision strategy used by the candidate, to name but a few. These forms of reliability are called test stability, and one way of obtaining this is by administering the same test or examination to the same group of individuals on two or more occasions and comparing the results. But this can only be an estimate, as the prior experience of having sat the same examination will tend to affect the second sitting in some way. In spite of this it is essential that we have some way of estimating stability effects for our assessments. Given all the possible sources of instability, we expect an upper limit of at most about 0.85 on the expected reliability of a multiple choice school examination. It is important to note I am not here trying to do full justice to issues of reliability, I am illustrating the importance of the application of psychometric principles.

Reliability in assessment is just the first step, however. A score can be perfectly reliable and still utterly invalid for a particular application. Astrological charts or diagnoses made on the basis of graphology (handwriting analysis) may be very reliable in that astrologers or graphologists using the same system will usually be in agreement about the significance of birth signs or of particular aspects of a person's handwriting. But this certainly does not mean that these techniques necessarily predict either personality or the future. This is assessed by the psychometric principle of validity. In order to assess validity we first need to be clear about the purpose of an assessment. There are various forms that validity can take, the primary ones being face validity, content validity, criterion-based validity and construct validity. Face validity is the extent to which an examination or test 'feels right' for the person taking it. In a personality test, for example, are the questions actually relevant to the stated purpose of the test? Or in an examination does the type of question reflect the social world of the candidates, or is it alien to them? Content validity can be demonstrated by matching the assessment specification to the curriculum that has been followed. If candidates are set examination questions that are outside the syllabus, then this represents a failure in content validity. Criterion related validity is important when a test or examination is used to predict future performance on some criterion. The validity of school examinations for university entrance, for example, can be assessed by following successful candidates throughout their university life. Construct validity attempts to analyse what it is that a test or examination is actually measuring. It requires many years of research in which underlying issues fundamental to a concept are addressed in many different ways and from many angles. Differences of approach concerning the curriculum, pedagogical method and assessment of key aspects of schooling such as learning the 'times tables' or the phonetic approach to reading illustrate the struggle to define the constructs of ability and achievement in mathematics and reading in a meaningful way.

The third psychometric principle we need to address is standardisation. There are, in essence, two forms of standardisation: norm-referencing and criterion referencing. In practice there is often a complex mix of the two in public systems, as Newton and Baird remind us. A norm-referenced test or examination compares the score of an individual with those of other candidates who took the test under similar circumstances. This group of candidates is called the norm group. The ultimate norm group would be the whole population of potential test takers. The standardisation of the WIAT (Wechsler Individual Achievement Test), for example, was carried out by obtaining a stratified random sample of 800 children throughout the UK. The proportion of children chosen mirrored those in the 2001 Census in terms of gender, ethnic group, locality and parent's educational level. The use of this type of referencing is important when educational psychologists need to match the achievement of an individual child against reasonable expectations based on how children of a similar age achieve in the population at large. Criterion referencing refers to a matching of a test score or an examination result to some objectively assessed reference point that details how a person with this score might be expected to perform in future training or in the workplace. Some have attempted to set these forms of standardisation against each other, but such arguments are usually vacuous as both are important in most real world settings, each in their own way.

The final psychometric principle is freedom from bias. Bias occurs when scores on a test vary depending on group membership. A test, examination or assessment procedure is said to be biased when its use results in adverse impact on one or more groups when compared with others. Groups can be defined in many ways, but becomes particularly significant in areas where anti-discrimination legislation is in force, such as gender, ethnicity, social circumstance, disability, sexual orientation and now, age. There are three principle types of bias: item bias, intrinsic test bias and extrinsic test bias. Item bias occurs when some items within a test show group differences that are disproportionate with the test as a whole. It might occur, for example, where a particular item contains English that is far too colloquial when addressed to candidates for whom English is not their first language. Item bias is, in principle, fairly easy to identify, but much more could be done to ensure that procedures are in place to keep it to a minimum. Intrinsic test bias occurs where a test or examination has differential reliability or validity for different groups, and much of the research on intrinsic test bias was associated with attempts to introduce positive discrimination policies, particularly in the US. But latterly there has been an increased recognition that, apart from item level bias, most of the bias found in assessment is extrinsic to the test or examination itself. More often, differences in test scores between groups come about as a result of the impact of real differences in society. Bias in, and the consequent adverse impact of, school examination results can to a large extent be accounted for by differences between localities in the quality of schooling, or of parental, peer and teacher expectation and support. These are themselves dependent on the impact of social policy on local demographics.

How do the psychometric principles relate to the evaluation and development of school examinations such as the A-level? Very much. First, we need to dispel a common myth that A-level results are judgements, not measurements, and hence escape the need to be judged by psychometrics. Judgements, as much as measurements, need to be reliable, valid, well standardised and free from bias. Hence the principles are unavoidable. Furthermore, the distinction is in many ways artificial. Psychometrics today is defined as the science of psychological assessment, not simply measurement, and this is important, particularly when an organisation has to decide on how an assessment is to be made. In recruitment, for example, it is not simply a question of whether to use psychometric tests, interviews, or other alternatives such as work sample exercises. Rather, it is a question of comparing the reliability, validity, standardisation procedure and extent of bias that lie in each, and deciding on the overall package. To do this common criteria are needed and these the psychometric principles supply.

Politics and misunderstandings abound in the school examinations domain, and application of the psychometric principles enables us to divide the fact from the rhetoric in the frequent debates that are now part of our lot. Given what we know about reliability, how has it come about that we experience demands that examination results must be completely reliable, something we know to be impossible? The misunderstanding arises because all tests and examinations inhabit a world of conflicting meanings and interpretations, and therefore need to be assessed in terms of their consequences as well as their psychometric characteristics. In education these include progression, educational treatment, admissions, setting and streaming. Outside education, tests not only assess, they also license, and once test results are used to license they cross a threshold that interfaces with the wider legal and social system of society at large. Hence the award, for example, of a driving licence, or of membership of the Royal College of Surgeons, or of a place at University, give legal entitlements that, while based on assessment, achieve a new status that are inevitably going to be the subject of controversy.

To validate a public examination, as with any other test, we need first of all to define its purpose. This is a basic requirement as otherwise we could not know whether its purpose was being met. It is a multifaceted issue as each examination serves a number of different purposes, hence a series of validations are required. Problems can arise if some of these purposes are in conflict. For example, we may find that increasing validity 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