link (i.e. authenticity, burden) and concerns about workload seem to be out-weighing possible advantages of coursework to validity in terms of construct representation (extrapolation link) and learning experiences (impact link). However, if the controlled assessments could maintain validity in terms of construct representation and learning experiences as well as reducing threats in relation to administration, then they could provide a more robust overall 'chain' of validity links.

References

Cronbach, L.J. (1988). Five perspectives on validity argument. In: H. Wainer & H.I. Braun (Eds.), *Test validity*. 3–17. Hillsdale, NJ: Erlbaum.

Cronbach, L.J. (1989). Construct validity after thirty years. In: R.L. Linn (Ed.), *Intelligence: measurement, the theory and public policy*. 147–171. Urbana: University of Ilinois Press.

Crooks, T.J., Kane, M.T. & Cohen, A.S. (1996). Threats to the valid use of assessments. *Assessment in Education: Principles, Policy and Practice*, **3**, 3, 265–286.

Cunningham, B. (1991). Coursework in GCSE English and English literature. In: SEAC (Ed.), *Coursework: learning from GCSE experience: an account of the proceedings of a SEAC conference*. London: Secondary Examinations and Assessment Council.

DfES (2005). *14–19 Education and Skills White Paper*. London: DfES.

Ellis, S.W. (1998). *Developing whole-school approaches to curriculum and assessment in secondary schools*. Paper presented at the British Educational Research Association Annual Conference, Queen's University, Belfast.

Frederiksen, J.R. & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher*, **18**, 9, 27–32.

Haertel, E. (1992). Performance measurement. In: *Encyclopedia of educational research*. 6th ed., 984–989. New York: Macmillan.

Harlen, W. (2004). A systematic review of the evidence of reliability and validity of assessment by teachers used for summative purposes. In: *Research Evidence in Education Library*. London: EPPI – Centre, Social Sciences Research Unit, Institute of Education.

Kingdon, M. & Stobart, G. (1988). *GCSE examined*. Lewes: Falmer Press.

Leonard, A. (1991). Case studies of school-based innovation. In: SEAC (Ed.), *Coursework: learning from GCSE experience: an account of the proceedings of a SEAC conference*. London: Secondary Examinations and Assessment Council.

Linn, R.L., Baker, E.L. & Dunbar, S.B. (1991). Complex, performance-based assessment: expectations and validation criteria. *Educational Researcher*, **20**, 8, 15–21.

Messick, S. (1989). Validity. In: R.L. Linn (Ed.), *Educational measurement*. 3rd ed., 13–103. New York: American Council on Education.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, **50**, 9, 741–749.

MORI. (2006). *Teachers' views on GCSE coursework: research study conducted for the QCA*. Available at: www.qca.org.uk/downloads/QCA-06-2735_MORI_report_teacher_views_coursework.pdf (accessed 13.10.06).

Ogborn, J. (1991). In-course assessment: creating and making best use of opportunities. In: SEAC (Ed.), *Coursework: learning from GCSE experience: an account of the proceedings of a SEAC conference*. London: Secondary Examinations and Assessment Council.

QCA. (2005). *A review of GCE and GCSE coursework arrangements*. London: Qualifications and Curriculum Authority.

QCA. (2006a). *A review of GCSE coursework. London: Qualifications and Curriculum Authority*. Available at: www.qca.org.uk/downloads/QCA-06-2736_GCSE_coursework_report-June-2006.pdf (accessed 13.10.06).

QCA. (2006b). *GCSE mathematics coursework: consultation summary*. London: Qualifications and Curriculum Authority. Available at: www.qca.org.uk/downloads/QCA-06-2737_maths_coursework.pdf (accessed 13.10.06).

QCA. (2007). *Controlled Assessments*. London: Qualifications and Curriculum Authority. Available at: www.qca.org.uk/qca_115533.aspx (accessed 06.08.07).

Scott, D. (1990). *Coursework and coursework assessment in the GCSE*. Coventry: University of Warwick, Centre for Educational Development, Appraisal and Research, CEDAR reports, no.6.

SEC. (1985). *Working paper 2: coursework assessment in GCSE*. London: Secondary Examinations Council.

SEC. (1986). *Working paper 3: policy and practice in school-based assessment*. London: Secondary Examinations Council.

Shepard, L.A. (1993). Evaluating test validity. *Review of Research in Education*, **19**, 405–450.

Taylor, M. (1992). *The reliability of judgements made by coursework assessors*. AEB Research Report RAC 577.

Wiliam, D. (1996). Standards in examinations: a matter of trust? *The Curriculum Journal*, **7**, 3, 293–306.

ASSESSMENT JUDGEMENTS

School-based assessment in international practice

Martin Johnson Research Division **and Newman Burdett** Cambridge International Examinations

Introduction

This article is based on research and experience in a wide variety of circumstances, educational back drops, social, cultural and political imperatives and, therefore, the proposals and guidelines need to be taken in context; it is impossible to argue whether a Ferrari or a Land Rover is a better car unless you know how it is to be used.

The term 'school-based assessment' (SBA) can conjure up diverse and not necessarily synonymous meanings which often include forms of ongoing and continual classroom assessment of a formative nature. Sometimes the term is simply used to distinguish localised assessment arrangements from other externally imposed forms of testing. In this article we have defined SBA in a more restricted sense; using it to describe the assessment of coursework. The UK Qualifications and Curriculum Authority (QCA) define coursework as 'any type of assessment of candidate performance made by the school or college in accordance with the specification (or syllabus) of the course of study that contributes to the final grade awarded for a qualification' (QCA, 2005, p.6). QCA go on to identify a number of activities that might be suitable for coursework assessment, and these include: written work and

extended essays; project work and investigations; practical experiments; production of works of art or other items; production of individual or group performance work; oral work or statistical and numerical tasks.

SBA can deliver strong educational benefits but like any powerful tool must be used with discrimination and care. SBA is a significant resource commitment, whether this burden lies with schools, education and curriculum bodies or assessment bodies, the resource implications need to be factored in and the benefits justified. SBA, like any educational assessment tool, must be fit for purpose and the analysis of whether it is successful can only be judged if the rationale for its introduction is clear. This article attempts to clarify how, and why, SBA has been successfully introduced in various contexts and the importance of the context in its success or otherwise.

Review of SBA research

Arguments are often framed in terms of the trade off between validity and reliability. Supporters of coursework suggest that it can increase examination validity, making what is important measurable rather than what is measurable important (SEC, 1985).

Despite this, Harlen (2004) cautions that evaluating one assessment method in terms of another, that is, evaluating coursework in terms of its reliability with timed, written examinations, can be problematic, overlooking the essential and important differences that might exist between them. Morrison *et al*. (2001) also suggest that such attempts lead to perceptions of a 'false equivalence', whereby both methods are understood to be equally effective at measuring the same skills, disregarding pedagogic imperatives.

What are the advantages of SBA?

One of the arguments often put forward for implementing SBA is that it reduces student anxiety which can have a significant impact on performance in written examinations (Lubbock and Moloney, 1984). This is particularly the case for tasks which are 'hard to get into' or depend heavily on insight.

Coursework can provide a wider range of evidence of candidates' achievements on different occasions, helping to ensure that the skills assessed reflect the wider curriculum. This could lead to a reduced emphasis on memory and a learner's ability to work quickly over a short period of time and a greater emphasis on research skills, interactive skills, motor skills, skills of adaptation and improvisation (Wood, 1991). Some skills and knowledge, especially those related to processes cannot be appropriately assessed in a terminal examination. It can also give pupils credit for initiating tasks and assuming responsibility for organising their own work. This also means that coursework assessment can correspond much more closely to the scale of values in the wider world, where the individual is judged as much by their style of working and ability to cooperate with colleagues as by the eventual product (SEC, 1985).

Coursework can provide the flexibility needed for assessment across a wide ability range through presenting pupils with tasks appropriate to their individual levels of ability. Research suggests that practical tasks set in authentic contexts can help less able students to understand 'what it is about' and to perform better (Gray and Sharp, 2001).

The 'assessment for learning' agenda rests firmly on the notion of giving clear learner feedback and encouragement. SBA allows teachers to capitalise on these formative qualities and promote achievement. Because of its proximity to the task, continual assessment can contribute to raising the quality of learners' work (SEC, 1985). Wood also highlights that coursework, above all other assessment techniques, is most likely to elicit 'positive achievement', focussing on what students know rather than what they don't know.

Why are there reservations about using SBA?

One of the most universally held perceptions (or misconceptions, depending upon viewpoint) is about lack of assessment reliability. Although acknowledging that the benefits of coursework generally outweigh any drawbacks, the QCA 2005 review of coursework identifies a number of concerns about the method, including the scope for plagiarism or other difficulties in work authentication. Whether or not it is a genuine concern, it can occupy a high public position and must be considered by policy makers and implementers. Additional workload, for both students and teachers, also features highly, especially amongst teachers where the burden of assessment can move from an external body (the exam board) to the teacher. This aspect was considered to be an issue of relevance for the *14–19 Education and Skills White Paper* published by the UK Deprtment for Education and Science which sought to review coursework arrangements 'to reduce the assessment burden' in some subjects (DfES, 2005, p.7). This raises issues about remuneration and resources; in a well designed SBA the process of teaching and assessment should be blurred and the overhead minimal.

Finally, there are issues of relevance; both to pedagogic methodologies and to learning outcomes. In many contexts, including the US, the UK, Australia, New Zealand, South Africa and Hong Kong amongst others, SBA has been proposed as a means of providing a more authentic assessment and educational experience, broadening the curriculum (Maxwell 2004), widening the range of syllabus outcomes assessed (Board of Studies NSW, 2003; Fung, 1998) and reducing the negative 'backwash' of summative assessment (Kennedy, 2006). But, as with any assessment tool, SBA can distort learning outcomes to meet the criteria, rather than the criteria reflecting learning outcomes. Similarly it has been accused of narrowing curricula and teaching to contexts that fit the criteria rather than contexts that enhance learning. In some subjects, most notably mathematics, the use of SBA as part of a generic educational policy has been argued to be at odds with competing teaching strategies, which might provide better educational outcomes (QCA, 2005).

What are the reported flaws of SBA?

Using generic criteria is often cited as a flaw in coursework implementation. The majority of GCSE coursework in the UK is based on generic rather than task-specific criteria, leading to inevitable inconsistencies in interpretation due to variances in teacher experience/expertise.

Beyond a teacher's individual interpretation of criteria, the concern of inappropriate teacher influence in coursework tasks is a key threat. There are suggestions that teachers can influence the organisation of portfolios in order to maximise student attainment. Although Wilmut *et al*., (1996) argue that there is a lack of research evidence about the possible nature and extent of bias in teacher assessment, it remains a high profile concern.

This also influences the debate over SBA's impact on standards and the generalisibility of teacher judgements beyond their immediate context. Accurate standardisation of grades over a large number of centres is difficult. Laming's (2004) psychological theories suggest that judgements are heavily influenced by context. Teacher judgements are prone to be influenced by the performances of students around them. Sadler (1998) reinforces this, suggesting that the use of an existentially determined baseline derived from how other students perform means that the teacher is unable to provide standards-oriented feedback because the judgements tend to norm-referencing. This cohort effect can also negatively impact on student 'ego' involvement. Where judgements are partly cohort-dependent students are more likely to interpret negative comments as being personal criticisms.

Context can also interfere with investigative assessment task design, and therefore inferences made about performance. Roberts and Gott (2004) suggest that a 'context effect' (the 'procedural complexity' or openness of a task) may necessitate the completion of up to 10 assessed investigations to be reasonably sure that the result was a reliable predictor of future ability. Rather than reducing assessment burden this might increase it.

The most damaging argument against the successful implementation of SBA is what is euphemistically termed 'construct irrelevant variance', or those factors that could be considered to give unfair advantage to some students (e.g. plagiarism, parental help given etc.).

What do empirical studies say about using SBA?

Wilmut *et al.* (1996) state that little has been published on the reliability of school-based assessments since a study which showed an average 0.83 correlation between schools' and an independent moderators' assessments (Hewitt, 1967). They go on to argue that this compares favourably with what might be expected from any two examiners marking an essay paper. They also suggest that Hewitt's findings are reinforced by those of Taylor (1992) who reported very creditable correlations (0.87–0.97) between pairs of moderators marking English and mathematics coursework folders.

Further research has reported that teachers are able to score hands-on science investigations and projects with high reliability using detailed scoring criteria (Frederiksen and White, 2004; Shavelson *et al*., 1992). Harlen refers to research suggesting the significance of assessment specification detail. Koretz *et al.* (1994) and Shapley and Bush (1999) report instances of poor assessment reliability where task specification was low.

Wood reports the findings of a study into coursework suggesting that coursework was a good discriminator in most of the subjects involved and not an easy source of marks (Stobart, 1988). Stobart explains that this was possibly because the assessments were collected over a longer period and contained more information to support discrimination between candidates.

Some studies suggest that assessment mode is a factor in the differential performance of boys and girls (Stobart *et al.*, 1992; Murphy, 1982; Newbould and Scanlon, 1981; Harding, 1980). These studies show that boys tend to be favoured by multiple choice questions and girls by essays and coursework, although Trew and Turner (1994) challenge such a conclusion, with Elwood (1995) suggesting that the effect of this on final grades is overstated.

When and why SBA should be used

SBA is arguably most effective as both an educational tool and as an assessment instrument when used to assess the acquisition of skills that are hard to demonstrate in a written examination (SEC, 1985, p.2; QCA, 2005, p.5). This applies especially to technical and creative subjects, science practical work and subjects where research or portfolio work would be naturally used in the course of teaching.

• Where it is a mechanism for achieving educational imperatives

Where skills are not being effectively taught in the classroom then, if appropriate, coursework can be used to ensure that skills *are* effectively taught. SBA can be a very powerful pedagogic device. Conversely, if poorly thought out it can have damaging consequences. This is clearly laid out in the Wilmut review:

> *If the primary goal is to maximise reliability then internal assessment might be an inappropriate tool. If the primary goal is to harness a powerful tool for learning then internal assessment may be essential.*

Cambridge International Examinations' (CIE) experience in implementing SBA systems around the globe suggests that it coincides with improvements in student performance. However, the untangling of cause and effect in these situations is very difficult. Implementing programmes where practical work has not previously been well taught requires a large input into teacher education and up-skilling to support effective assessment. It is possible that this, rather than effects of the SBA, contributes to observed improvements. Either way such improvements are a positive benefit of the introduction of SBA and it provides a framework and feedback mechanism to maintain improved standards, both for students and for the teachers as learners. In Botswana, students in trial schools implementing SBA showed not only an improvement in practical performance but also an improvement in their understanding and knowledge of the subject as a whole as judged by performance on written papers (Thomas, 2006). Again this needs further research to determine if this is a direct benefit of the pedagogy of SBA or whether implementing SBA encouraged teachers to become more reflective of their teaching methods and therefore better pedagogues. Similarly, in international cohorts where SBA is offered as a choice to other forms of examination, student performance is better on *objective* tests. Again this finding does not distinguish cause and effect; teachers opting to use SBA are likely to do so for educational reasons and are potentially more likely to be teachers with (or in schools with) a stronger educational philosophy than those choosing other forms of assessment. The flip side of this is that when SBA is first implemented to improve teaching, an increase in standards over the first few sessions is expected and standard setting should be criteria-based, avoiding statistical moderation unless there is clear evidence that the criteria are being mis-applied; this is best dealt with by centre moderation and teacher training.

• Where it offers improved validity and more focussed and efficient assessment

SBA appears to be commonly used in practical and applied subjects which try to capture process skills. This is not an unsurprising finding as these skills can be difficult to accurately assess in a terminal written

assessment and attempts to do so might distort validity and have an adverse wash-back effect on teaching and curriculum. Well structured SBA can lead to good formative assessment and a summative outcome and have a beneficial effect on student learning without sacrificing reliability, discrimination or validity.

• Where there is a desire to create more active learners, improve teacher feedback or implement specific pedagogic strategies

Some teaching approaches, by their very nature, can only be implemented if there is teacher assessment as part of the learning cycle. If teachers are unwilling or unsure how to implement these strategies then externally imposed SBA can be considered. It is important to realise, however, that high-stakes SBA is not necessary for this and might have a negative wash back effect. This has been seen in UK science SBA practice, where requirements to ensure reliability and differentiation have led many teachers to claim that they undermine the benefits and lead to a narrowing and stagnation of teaching of practical work; accusations of 'hoop jumping' are not uncommon (QCA, 2005, p.10).

Feedback on performance is vital for any learner to improve their learning and guide their future learning. This applies equally to teachers, who in order to improve their teaching need to practice what they preach and become reflective learners. This feedback is a vital part of the algorithm. Indeed, this is one of the strands that was criticised in the QCA review of UK SBA practice (QCA, 2005).

When and why SBA should not be used

• To promote good teaching when SBA does not fit comfortably into the subject area

Assessment and curriculum are closely linked but assessment should encourage and support the curriculum. Assessment should reinforce good teaching; but it can be a crude tool and can prove counterproductive if used carelessly. Teachers should be encouraged to teach well by ensuring that assessments reward the learning outcomes defined in the curriculum, and those learning outcomes should reflect good pedagogy. If a curriculum is to be encouraged away from a transmissive pedagogy to produce more inquiring students, focussing on skills and the application of knowledge, SBA by itself will not deliver this change. Experience suggests teachers and students are very efficient at subverting SBA to provide a line of least resistance in terms of withstanding change. Teachers tend to maintain their more familiar didactic pedagogy, using common strategies such as the 'over-structuring' of tasks, coaching, exemplars and re-use of formulaic tasks in order to meet the requirements of the assessment without students necessarily fully engaging with the learning outcomes. Similarly, if the subject does not readily lend itself to SBA then it is unlikely to be successful; the strong trend in the UK to move from mathematics specifications with compulsory course-work to those without would indicate that this is a subject where the benefits of SBA are seen by the teacher to be minimal and out-weighed by the detrimental aspects. Evidence from the QCA review (2005) found that 66% of mathematics teachers indicated that coursework was sometimes problematic compared with largely positive reflections from English, history, psychology, geography and design technology teachers.

• When external pressure for reliability places a large burden on teachers and assessment bodies

SBA will always be open to accusations of bias because of the close relationship between teacher and student and the potential vested interests in improving the outcome. It is noticeable that in situations where SBA works well there is usually a lack of intense student competition (e.g. in countries where funded University attendance is guaranteed) and where teacher performance is not judged by student performance. Recent reports from Sweden highlight increased concerns about grade inflation when teacher performance management systems and national educational auditing are linked to student performance (Wikström and Wikström, 2005). Similarly, pressure to perform well in league tables has been cited in the UK as a distorting influence on the success of SBA, with Wilmut *et al*. highlighting this as one of the situations where SBA should be avoided.

• Where SBA is dissociated from pedagogic principles and hinders learning feedback mechanisms

Along with assessment validity, educational validity is a key reason to introduce SBA. Without either of these *raisons d'être* SBA, as with any assessment tool, should not be employed; other forms of assessment will be more productive and more supportive of good learning.

Some commonly encountered problems

Experience in international contexts suggests that the following problems occur regularly in discussions between assessment and education professionals who have recently implemented coursework into curricula.

• The focus is on assessment not learning

The only solution to this is to restructure the scheme of assessment to encourage good learning and ensure that even if teachers teach to the test that students benefit from the learning experience. Assessment that involves clearly communicated learning targets for the students with a format that encourages active learning can help to make assessment and learning mutually supportive.

• Teachers continue to use a transmissive pedagogy leading to students focussing on learning rote responses rather than transferable skills

Again, pragmatic use of assessment methods and employing a mixed portfolio of assessment tools can make this a technique of diminishing returns. If students can be encouraged by the way that the SBA system is implemented, then they might avoid such short-sighted techniques realising that they will be disadvantaging themselves, both in terms of learning and scores. Overly prescriptive and restrictive criteria can lead to this and should be reviewed.

• Teachers find it difficult to make accurate assessments of students' work

One of the main focuses of many SBA reviews is the issue of reliability. Wilmut *et al*. make two helpful observations on this:

> *Reliable assessment needs protected time for teachers to meet and to take advantage of the support that others can give.*
>
> *Teachers who have participated in developing criteria are able to use them reliably in rating students' work.*

Neither of these may be practicable in all situations but should be kept in mind when developing systems. Simple and clearly expressed criteria relating to clear learning outcomes can also help, along with the avoidance of over-reliance on vague adjectives and other subjective terms in the criteria.

• Narrowing educational outcomes

SBA should be flexible enough to encourage teachers and students to explore the boundaries of the curriculum. It is often the case that an emphasis on reliability rather than validity can lead to SBA encouraging a conservative approach to interpreting the curriculum. If we try to avoid conflating different skills which may be mutually dependent on each other for a successful performance, this can help clarify to the learner what is required both in terms of individual skills and how they then link together. As in any assessment, we need to think clearly about the strategies learners will employ in responding to an assessment task. Using appropriate assessment can encourage the scaffolding of learning by making clear the stages in a task to both the learner and assessor. This can also facilitate the identification of a learner's potential weaknesses or misconceptions. This clarity might also help to create the confidence to explore the curriculum more widely by encouraging a more holistic view of learning.

• SBA leads to disinterest and low morale

This can apply to both educators and students. This is a symptom of a variety of the aspects already described. A well-designed SBA system should encourage good education, part of which is to instil a sense of enquiry into students. If it is not doing this then it needs to be reviewed. Involving students in the learning process, ensuring the SBA allows for a constructive feedback loop with the student, and making students aware of the learning outcomes they are aiming for, can all help and should be considered when designing SBA. Similarly, the system should allow for flexibility and individual learning progression.

Summary

It is important to highlight that none of the above findings or recommendations should be taken in isolation and many can apply equally to other forms of assessment. It is also the case that, arguably more than in any other kind of assessment, SBA entangles pedagogy and assessment issues such that one cannot be considered separately from the other. Public and professional perception that coursework was increasing student and teacher workload without a perceived increase in educational benefits led the UK QCA to conduct a review of coursework at GCSE level. This review highlighted several recommendations and these prove universally applicable to any SBA. These include the following advice:

- There is a need to have mechanisms in place to avoid malpractice, including the need for clear roles, responsibilities and constraints on teachers and parents in relation to coursework.
- It is also necessary to have effective mechanisms for standardisation of assessors.
- There is a need for a clearly defined purpose and format for feedback.
- It is important to decide whether SBA is a *necessary* and *appropriate assessment* instrument for specific subject learning objectives. (QCA, 2005, p.22)

It is essential to remember that any assessment or educational reforms require the support and participation of the stakeholders; due to its high visibility in the daily lives of students, the introduction of SBA often requires that this support be even more positive.

References

Board of Studies NSW (2003). *HSC assessment in a standards-referenced framework – A Guide to Best Practice*. State of New South Wales, Australia: Board of Studies NSW.

DfES (2005). *14–19 Education and Skills White Paper.* London: HMSO.

Elwood, J. (1995). Undermining gender stereotypes: examination and coursework performance in the UK at 16. *Assessment in Education*, **2**, 3, 282–303.

Frederiksen, J., & White, B. (2004). Designing assessment for instruction and accountability: an application of validity theory to assessing scientific inquiry. In: M. Wilson (Ed.), *Towards coherence between classroom assessment and accountability, 103rd Yearbook of the National Society for the Study of Education part II*. Chicago, IL, USA: National Society for the Study of Education.

Fung, A. (1998). *Review of public examination system in Hong Kong: Final report*. Hong Kong: Hong Kong Examinations Authority.

Gray, D. & Sharp, B. (2001). Mode of assessment and its effect on children's performance in science. *Evaluation and Research in Education*, **15**, 2, 55–68.

Harding, J. (1980). Sex differences in performance in science examinations. In: R. Deem (Ed.), *Schooling for Women's Work*. London: Routledge & Kegan Paul.

Harlen, W. (2004). *A systematic review of the evidence of reliability and validity of assessment by teachers used for summative purposes*. Eppi Review, March 2004.

Hewitt, E. A. (1967). *The Reliability of GCE O Level Examinations in English Language*. JMB Occasional Publications 27. Manchester: Joint Matriculation Board.

Kennedy, K. J., Chan, J. K. S., Yu, F. W. M. & Fok, P. K. (2006). *Assessment of productive learning: forms of assessment and their potential to enhance learning*. Paper presented at the 32nd Annual Conference of the International Association for Educational Assessment, Singapore.

Koretz, D., Stecher, B. M., Klein, S. P. & McCaffrey, D. (1994). The Vermont Portfolio Assessment Program: findings and implications. *Educational Measurement: Issues and Practice* **13**, 5–16.

Laming, D. (2004). *Human judgment: The eye of the beholder*. London: Thomson Learning.

Lubbock, J. and Moloney, K. (1984). *Coursework Assessment*. London: City and Guilds.

Maxwell, G. S. (2004). *Progressive assessment for learning and certification: some lessons from school-based assessment in Queensland*. Paper presented at the 3rd Conference of the Association of Commonwealth Examination and Assessment Boards, Fiji.

Morrison, H., Cowan, P. & D'Arcy, J. (2001). How defensible are current trends in GCSE mathematics to replace teacher-assessed coursework by examinations? *Evaluation and Research in Education*, **15**, 1, 33–50.

Murphy, R. J. L. (1982). Sex differences in objective test performance. *British Journal of Educational Psychology*, **52**, 213–19.

Newbould, C. & Scanlon, L. (1981). *An analysis of interaction between sex of candidate and other factors*. TDRU: Cambridge.

QCA (2005). *A review of GCE and GCSE coursework arrangements*. London: Qualifications and Curriculum Authority.

Roberts, R. & Gott, R. (2004). Assessment of Sc1: Alternatives to coursework. *School Science Review*, **85**, 313, 103–108.

Sadler, R. (1998). Formative Assessment. *Assessment in Education*, **5**, 1, 77–84.

Secondary Examinations Council (1985). *Working Paper 2: Coursework assessment in GCSE*. London: SEC.

Shapley, K. S. & Bush, M. J. (1999). Developing a valid and reliable portfolio assessment in the primary grades: building on practical experience. *Applied Measurement in Education*, **12**, 11–32.

Shavelson, R. J., Baxter, G. P. & Pine, J. (1992). Performance assessments: political rhetoric and measurement reality. *Educational Researcher*, **21**, 22–27.

Stobart, G. (1988). *Differentiation in practice: A summary report*. London: University of London Schools Examination Board.

Stobart, G., Elwood, J. & Quinlan, J. (1992). Gender bias in examinations: how equal are the opportunities? *British Educational Research Journal*, **18**, 3, 261–276.

Taylor, M. (1992). *The reliability of judgements made by coursework assessors*. AEB Research Report RAC 577.

Thomas, I. (2006). *Internal Grading Report*. Cambridge: Cambridge International Examinations.

Trew, K. & Turner, I. (1994). Gender and objective test performance. In: K. Trew, G. Mulhern & P. Montgomery (1994). *Girls and women in education*. Leicester: The British Psychological Society, Northern Ireland Branch.

Wikström, C. & Wikström, M. (2005). Grade inflation and school competition: an empirical analysis based on the Swedish upper school grades. *Economics of Education Review*, **24**, 3, 309–322.

Wilmut, J., Wood, R. & Murphy, R. (1996). *A review of research into the reliability of examinations*. Nottingham: University of Nottingham, School of Education.

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

Using simulated data to model the effect of inter-marker correlation on classification consistency

Tim Gill and Tom Bramley Research Division

Introduction

Measurement error in classical test theory is defined as the difference between a candidate's observed score on a test and his or her 'true' score, where the true score can be thought of as the average of all observed scores over an infinite number of testings (Lord and Novick, 1968). The observed score X on any test is thus:

$$X = T + E$$

where T is the true score and E the (random) error component. Whilst classical test theory recognises several sources of this measurement error, arguably the source of most concern to an awarding body is that due to markers – in other words the question 'what is the chance that a candidate would get a different mark (or grade) if their script were marked by a different marker?' (Bramley, 2007). Therefore, for the purposes of this article, the E in the above equation refers to marker error only. Other factors affecting measurement error such as the candidate's state of mind on the day of the exam or whether the questions they have revised 'come up' may be thought of as more acceptable by the general public; these are considered to be the luck of the draw. Getting a different grade dependent on the marker is much harder to accept.

However, the marking of exam papers is never going to be 100% reliable unless all exams consist entirely of multiple-choice or other completely objective questions. Different opinions on the quality of the work, different interpretations of the mark schemes, misunderstandings of mark schemes, or incorrect addition of marks all create the potential for candidates to receive a different mark depending on which examiner marks their paper. Awarding bodies put great effort into annual attempts to increase reliability of marking with standardisation meetings, scrutiny of sample scripts from each marker and scaling of some markers. However, these measures are far from perfect: examiners may make different errors in the scripts that are sampled than in other scripts. Scaling is a broad-brush approach, and it has been shown that it can cause more than 40% of the marks given by the scaled examiner to be taken further away from the 'correct' mark (Murphy, 1977 quoted in Newton, 1996).

Arguably, however, the real concern for examinees is not that they might get a different mark from a different examiner, but that they might be awarded a different *grade*. Investigations of the extent to which this occurs have been relatively few, judging by the published UK research literature (see next section for a review), probably because of the cost associated with organising a blind double-marking exercise large enough to answer some of the key questions. The purpose of this study was to use *simulated* data to estimate the extent to which examinees might get a different grade for i) different levels of correlation between markers and ii) for different grade bandwidths.

To do this we simulated sets of test scores in a range of scenarios representing different degrees of correlation between two hypothetical markers, and calculated the proportion of cases which received the same grade, which differed by one grade, two grades, etc. The effect of grade bandwidth on these proportions was investigated. Score distributions in different subjects were simulated by using reasonable values for mean and standard deviation and plausible inter-marker correlations based on previous research. The relative effect on unit grade and syllabus grade was also investigated.

Correlation is traditionally used as the index of marker reliability. Here we discuss some other indices and explore different ways of presenting marker agreement data for best possible communication.

Background and context

It is important at this point to emphasise a distinction that comes up in the literature on misclassification in tests and exams. This is the difference between classification *accuracy* and classification *consistency*. 'Accuracy' refers to the extent to which the classification generated by