Issue 5 January 2008

Research Matters

CAMBRIDGE ASSESSMENT

University of Cambridge
Local Examinations Syndicate

150 YEARS 1858–2008

If you would like to comment on any of the articles in this issue, please contact Sylvia Green.
Email:
researchprogrammes@cambridgeassessment.org.uk

The full issue and previous issues are available on our website:
www.cambridgeassessment.org.uk/ca/Our_Services/Research

Research Matters : 5

A CAMBRIDGE ASSESSMENT PUBLICATION

Foreword

This issue of *Research Matters* is testimony to the diversity of the education system in England – not 'social diversity', but diversity in assessment and qualifications. Andy Green, in his seminal book *Education and State Formation* (1990) compared England, Germany, France and the USA as a means of understanding why the English education is so diverse in its form and content. The influence of diverse lobby groups has been historically strong and remains so; freedoms for local systems and policies to condition provision have been sustained despite increasing centralised control over curriculum and related assessment; discontinuities in phases of education remain; and continued failure to establish credible vocational options has led to unclear messages to young people about routes through education and training. The 'melting pot' characteristic of the system, combined with constant 'pendulum swing' changes, has meant the emergence of highly diverse assessment practice and the failure of any one paradigm to dominate. Extracting clear messages about the direction in which we should take both assessment policy and technical development is thus a complex business. Understanding the history of assessment, drilling down into different types of assessment and measurement models, and understanding the outcomes data, all play a role; the articles in this issue take a look at all of these themes.

Tim Oates *Group Director, Assessment Research and Development*

Editorial

The first article is a presentation given by Andrew Watts as part of the programme for the 2007 Annual Alumni Weekend of the University of Cambridge. He describes how public examinations were first introduced in England for secondary age students and charts the development of the system.

A main theme of this issue is the psychology of assessment and the way that judgements are made. Suto, Crisp and Greatorex have worked on a number of linked research studies considering the marking process from a range of perspectives related to human judgement and decision making. In their article they provide an overview of their research in the context of GCSEs and A-levels. Nádas and Suto continue the theme in their article on self-confidence and insight into marking accuracy among GCSE maths and physics markers. Novakovič discusses assessment judgements made in Angoff awarding meetings and the impact that the provision of performance data has on decision making.

Victoria Crisp reviews research literature related to coursework at GCSE and also outlines potential changes to the way coursework may be included in future examinations. Johnson and Burdett continue the discussion in the context of school-based assessment in international practice. They consider when and why school-based assessment should or should not be implemented and how potential problems could be addressed. Two articles focus on examinations research, one in relation to classification consistency and the other to the patterns of GCSE and A-level uptake. Gill and Bramley report on the use of simulated data to model the effect of inter-rater correlation on classification consistency. Although this research is based on simulated data, it does add to the current debate on the need for education about measurement error. The second article highlights two new statistical reports available on the Cambridge Assessment web-site.

The final article is from Elizabeth Gray of OCR's operational research team. It provides an overview of the kind of work carried out by the team and some of the issues that they need to address in the context of quality and standards.

Sylvia Green *Director of Research*

Independent examination boards and the start of a national assessment system

Andrew Watts Cambridge Assessment Network

Examinations in the Age of Reform

The 1850s was the decade in which public examinations were run for the first time for secondary age students in England. To describe what came into being at first as 'a system' would be to misrepresent the ad hoc and small-scale nature of the operation. But the methods of running public examinations chosen in those early days has remained with us for 150 years. The effort to get something started was made by committed individuals and, with minimal government help, they enlisted the support of independent institutions. Indeed, at the beginning of that century the country did not expect the government to take responsibility for the education of its citizens, let alone for a system of public examinations. However, reform was in the air.

The introduction of examinations to the universities at the beginning of the century was seen as having been successful in raising standards at the universities. Students were now being required to demonstrate that they deserved their degrees and university posts were offered on the basis of merit and not of patronage. Thus people who had been through the universities began to argue that exams would be a tool for improvement on a wider scale (Montgomery, 1965). Under the influence of thinkers like Jeremy Bentham and John Stuart Mill, the idea took root that it was the responsibility of government to stimulate society to improve itself. The aim was, according to Mill's familiar phrase about Bentham's 'utilitarian' doctrine, the greatest happiness of the greatest number, and the engine which would drive this movement would be 'self-interest', or to use a less pejorative phrase, 'self-help'. As part of his plan for a modern society, Bentham (1827) worked out an elaborate examination system for applicants to the Civil Service and Mill (1859) later proposed a system of compulsory education based on an examination system.

Teacher accreditation schemes

As well as in the universities, public examinations were established in other areas before they were established for school students. In 1846 a scheme was set up by the government under which pupil-teachers were apprenticed to schoolmasters and mistresses. They started at the age of 13 and were examined at end of the fifth year of their apprenticeship by Her Majesty's Inspectorate (HMI), which had been set up in 1839. Prizes and scholarships were awarded to successful candidates. In addition to this scheme, grants were also made to students at training colleges. Both of these processes were managed through annual exams administered by HMI. For the colleges, inspectors at first paid individual, annual visits but from 1846 common exams were set simultaneously for all. As Roach (1971) has pointed out, the examination system for teachers was the first common test in England set on a general syllabus and taken in a number of separate places.

In addition to this government-run scheme, examinations for teachers were set by The College of Preceptors, which was set up in 1846. This independent organisation was given a royal charter in 1849 with the stated aims of improving teachers' knowledge and of granting certificates of fitness for office to teachers. College of Preceptors members were those working in private schools run for 'middle class' pupils. (For 'middle class' schools, see below.) Very few teachers came forward to take these exams, but in 1850 the College started examining pupils and these exams were fully operational in 1854. Again Roach (1971) points out that these were the first external examinations held for students in middle class schools in the country. However, the College of Preceptors was a small organisation and it lacked prestige, particularly as the public were suspicious of exams that were run by teachers (Montgomery, 1965).

Society of Arts[1]: Trade exams

From the 1820s onwards, those who had reached their teens with or without elementary schooling, could attend a growing number of private educational institutes, called 'Mechanics Institutes'. One of these was started in Wandsworth in 1853 by James Booth, whose important role in setting up the Society of Arts exams is described by Frank Foden (1989). At Wandsworth the Institute catered for 'the instruction of the children of artisans and small tradesmen in the knowledge of common things, that may be turned to practical usefulness in after life'. For a charge of one shilling or 1/6d a week children were taught 'a little of mechanics, chemistry, use of the steam engine, and geography, history and arithmetic and their bearing in relation to trade' (Foden, 1989).

In 1853 the Society of Arts (SA) proposed 'a scheme for examining and granting certificates to the class students of Institutes in union with the society [of arts]'. The first exam was offered in 1855 but only one candidate applied. (He was a chimney sweep called William Medcraft, and he studied at the Belmont Mutual Improvement Society.) Exams were offered again by the SA in 1856, and this time 42 candidates presented themselves. (This time William Medcraft managed to obtain pass certificates in arithmetic, algebra and geometry.) Foden credits James Booth with much of the work on rescuing the SA examinations, for he became chairman of the SA Board of Examiners after the debacle of 1855, and it was his revised scheme that can be called the blueprint for all future schools examinations. (See Foden, 1989, chapter 8 for the background to this and the next section.)

The Society of Arts fairly quickly, in 1859, handed over its examining activity in the Sciences to a government body, the Department of Science and Art. This department, one of whose aims was to encourage schools to take on the teaching of Science, distributed money to schools

1 The Society of Arts (which became the Royal Society of Arts in 1907), had been founded in 1754 with the aim of promoting art, industry, commerce and invention by awarding prizes and grants. It promoted the Great Exhibition of 1851 and at that time Prince Albert was its President.

on the basis of their pupils' performance in exams. Later (in 1879) the SA handed its exams in technology to the City and Guilds of the London Institute, thus keeping only its commercial exams, which RSA Examinations retains to today. The Society of Arts exams made, and have continued to make, a very significant contribution to the development of adult and work-orientated education in the country.

A blueprint for an exam system

As well as their impact on adult education, the SA's exams had a significant effect on the development of school exam administration, for they had demonstrated what could be made to work in practice. The following were features of the SA system, set up by the committee chaired by James Booth, and these features were taken on by other agencies, thus becoming part of the examination system with which we are still familiar.

Exams were to be held annually (in March) and exam timetables were published in advance.[2] Society of Arts rules allowed single subject entry and it was assumed students would take up to three subjects. (The university Locals later required students to attempt a full range of subjects.[3]) Sample question papers were sent to institutions and candidates were encouraged not to spend time on subjects that would not come up in the papers (Foden, 1989). Soon this led to the production of examination syllabi, which came to dominate the teaching syllabus (as indeed was hoped by the reformers). Question papers were set by the examiners and quickly they required only written answers. Initially, the invigilation was conducted by the examiners themselves, which enabled oral examinations to be conducted at the discretion of the examiner, usually to confirm the award of an 'excellent' grade. This was the model that James Booth knew from his days at Trinity College, Dublin and it presupposes a personal link between the examiner and the candidate which was quickly lost. The format of timed, essay-type questions was not old: it was the product of the constraints and conditions under which public examining began, and was an inevitable product of the increasing number of candidates.

The examiners who marked the papers were regarded as the 'best authority' and were chosen for their eminence in their field. It was not until much later in the century that teachers began to be involved in external examining. (The SA's 43 examiners in 1856 included: 14 Fellows of the Royal Society, 13 professors – 8 from King's College, London, The Astronomer Royal, Two Reverend Principals of Teacher Training Colleges, 2 HMI, including Frederick Temple, a future Archbishop of Canterbury.) Students' performances were divided into class 1 (excellence), 2 (proficiency), 3 (competence). Those who didn't make the grade were 'refused' or more colloquially, 'plucked'. No credit was given for a smattering of knowledge, poor spelling, or illegibility. The 1st class was very cautiously awarded.

It was felt from the start that feedback to centres should be given after the exams. This was of particular significance because a central purpose was to encourage improvements in schools and institutes. The examiners therefore took on an authoritative tone in their reports and saw it as their business to point out the deficiencies not only of the candidates but of the teaching programmes which had produced them.

'Middle class' schools

One familiar aspect of the Victorian era is their openly accepted division of English society into classes. The rapidly expanding middle class was seen to include a range of people from newly-successful industrialists to clerks and book-keepers, farmers, shop keepers and tradesmen. The issue of the need for examinations was discussed in class terms, as the need to improve 'middle class' secondary schools was seen as a major issue. Middle class schools did not attract government support, they were privately run and parents paid fees. Some of the schools were long-established grammar schools which had been allowed to deteriorate, some had historic endowments to provide education for certain groups of children and some were institutions started in people's private property and run by them and their families. Some of these private schools were good, but in many cases they were not. Charles Dickens' portrayal of schools in *Nicholas Nickleby*, *David Copperfield* and *Our Mutual Friend* gives a campaigning novelist's view of early Victorian middle class education and teachers.

Many Victorian reformers focussed their attention on this class of pupils and their schools, none more thoroughly than Nathaniel Woodard. He described the range of middle class people by dividing them into two grades: 1- gentlemen of small incomes, solicitors, surgeons, unbeneficed clergy, army and navy officers; 2- trades people, retail shop owners, publicans, 'respectable tradesmen dealing with the higher classes' (Woodard, 1848). Woodard was concerned that the church should lead the way in providing schooling for the children of such groups and he set about raising funds and then founding private schools, which are still known as 'Woodard Schools' (Woodard, 1851). He responded to the needs of the middle classes by founding schools: others envisaged public examinations as the way to improve middle class education.

The Oxford Local Exams

The start of the Society of Arts exams was significant not only for its own sake, but also because that experience played a part in the establishment of examining at Oxford and Cambridge. Two people who were involved in both the SA's and Oxford's exams were Viscount Ebrington and Frederick Temple. Ebrington was a land owner in Devon, and an MP, and he became Chairman of the Society of Arts' Council in 1854. This was the year in which the first proposal to set up an exam system was put to the SA. His particular concern was for the education of farmers, labourers and country tradesmen – a country equivalent of a wider concern for the poor standards of education for 'middle class' children. Ebrington's plan was for setting up a system of 'county degrees, awarded on the basis of county examinations'. Finding SA reluctant[4] he and another local landowner set up their own exam in Exeter in Easter, 1856 with support of the Bath and West Society. Ebrington offered a prize of £20 for the best performance of any young man between 18 and 23 who was the son or relative of a Devon farmer.

2 It is interesting to note that advances in technology, such as steam printing and the introduction of a postal service, made such an enterprise possible. Also, when local exams were planned, the examiners were able to reach local centres by railway.

3 The Locals and later the School Cert., in 1918, fostered the system of requiring students to sit for groupings of subjects, but in 1951 the 'O' level system went back to single subject entry.

4 Ebrington also thought that SA had insufficiently strong influence on public opinion to hold examinations. It was also the case that the SA planned its exams for adult workers not school students.

The Exeter committee asked the Department of Education for help in setting up their local exam and they were given the help of Frederick Temple[5]. It was Temple who became the primary link with Oxford University and in February 1857 Temple wrote to Robert Scott, Master of Balliol College proposing that the university should run local examinations. He drew up a clear and detailed scheme showing how it would work (Sandford, 1906, quoted in Roach, 1971). Roach concludes that it was Temple who was responsible for setting out a practical scheme of examining which convinced Oxford University that they could run such a system.

In 1857 a committee set up by the Oxford University Senate worked on Temple's idea. The scheme received a warm welcome. The *English Journal of Education* (1857) wrote that Temple 'had struck the key to a thousand hearts'. In June 1857 the University of Oxford Delegacy of Local Examinations (UODLE) was established by statute. Its aim was to conduct examinations of non-members of the university. This was widely seen as part of the movement to reform universities and make them become more democratic and socially involved. At a celebratory meeting in Exeter, Temple stated: 'The universities should be made to feel that they have an interest in the education of all England' (Sandford, 1906). The first Oxford Local Examinations were conducted in the summer of 1858 in 11 centres.

The Cambridge Local Exams

In Spring 1857, Cambridge University received a deputation from Birmingham and memorials from schools in Cheltenham, Leeds and Liverpool requesting that the issue of offering local exams be considered. The Council of Senate recommended that a syndicate be set up. This was done on 4th June and the syndicate reported on 19th November. It proposed:

- Examinations for pupils under 15 (16 was eventually agreed) and 18.
- The subjects to be examined were English Language and Literature, History, Geography, French, German, Latin, Arithmetic, Mathematics, Natural Philosophy, Religious Knowledge (unless parents objected).
- An award of 'Associate of Arts' to successful senior candidates. This proposal, accepted at Oxford, was dropped at Cambridge after intense debate.

The University of Cambridge Local Examinations Syndicate (UCLES) was eventually set up in February 1858, and the first examinations were held in December 1858. The papers were set and marked by members of the university's teaching staff, and administered by them in each locality with the support of 'local committees'.[6] It was hoped that eventually Oxford and Cambridge would work together, possibly by running exams in alternate years,[7] but this did not come about.

5 Frederick Temple was brought up in Devon, went to Oxford, and became a fellow of Balliol. Then to Education Department. Served as Principal of Kneller Hall, the training college for workhouse school masters. Then in 1855 became inspector (HMI) of men's training colleges. In November 1857 became Head of Rugby School and later was Archbishop of Canterbury. Temple was involved in SA exams as Examiner in English and was an influential member of SA's Board.

6 This link accounts for the word 'local' in UCLES' and UODLE's names. The SA exams were at first administered only in London. Candidates had to travel to SA headquarters to take the exams, a considerable disincentive. So there was quickly a demand for 'local' examinations in candidates' home towns. In the second year of its exams the SA ran a 'local' centre in Huddersfield. Even so, the expense continued to be a problem for most students at the time.

7 At first Oxford ran its exams in June and Cambridge in December. In 1860 the suggestion was revived that they should share the running of the exams by having them in alternate years. Agreement on this could not be reached and they continued on their separate ways.

Examinations and girls' education

An area in which the hopes of the reformers were fulfilled, though they had not made this area a main focus of their plans, was the effect on girls' education of the introduction of public examinations. Emily Davies, eventual founder of Girton College, agitated for girls to be able to enter the Cambridge Locals (Stephen, 1927, quoted in Roach, 1971). She wanted girls to be judged on the same standards and curriculum as boys and definitely turned down any idea that girls should follow separate syllabuses or tackle different exams (which was all that was offered by Oxford initially). At first the Cambridge Syndicate only agreed that girls could 'unofficially' sit the exams to be taken in December, 1863. This gave the girls only 6 weeks to prepare, but the campaigners were determined to make the best of the opportunity. 83 girls took the exams in London. Miss Buss, Head of North London Collegiate School for Ladies, who was strongly in favour of exams for girls, entered 25 candidates for the UCLES' experiment in 1863.

The next step was a memorial in 1864 to the Cambridge Vice Chancellor that girls should be able to enter the Cambridge Locals officially. It contained nearly a thousand signatures. A positive report was published in February 1865 and, by a narrow majority in the Senate, entry for girls on the same basis as boys was agreed for a three year period.[8] In 1865, 126 girls took exams in London, Cambridge, Brighton, Manchester, Bristol, Sheffield. In 1867 entry to Cambridge Locals was made permanent for girls.

Miss Buss spoke at her school prize giving in 1871 of the Locals' good effect, saying, 'There can be little doubt as to the good effect of these examinations on girls' education.' In 1891 she told the governors of her school: 'Our practice has been to use the public examinations of the Cambridge and London Universities for the purpose of school external examination … Since our scheme was passed, nothing less than a revolution in the education of girls and women has taken place' (Headmistress's Reports to Governors, 1971). Roach (1971) concludes that the Locals were 'one of the most important levers in raising the whole level of women's education throughout the country.'

Criticisms of examinations

There was satisfaction within the examination boards that their efforts were indeed helping to raise the standards of secondary education generally. At the same time, however, criticisms of examinations were forcefully aired and the points made are those with which a modern reader will be familiar. One criticism commonly reported in the reports of HMI was the domination of schools' curricula by examinations, which led schools to provide too narrow a range of study, with any idea of a broader education being discouraged.

A German observer of the English education system (Weise, 1877) commented on the irony that schools were so keen for examinations over which they had so little control. Such criticisms were aimed specifically at the Locals, which appointed university men as examiners and in which teachers claimed the standards were set too high and the examiners were out of touch with schools.

8 The discussion of girls' involvement in exams revealed some deeply held views about what girls could and could not do. One Headmistress in a school in Malvern was reported as saying that 'no girl ought to go through anything that was public.' (Transactions of the Social Science Association, 1872)

Cramming and too much competition were claimed to be causing pupils to become over-strained. The argument was put forward that examinations were damaging the students' health[9], an argument that was particularly advanced when the examining of girls was under discussion. At a meeting in November 1871 the chairman of the Social Science Association, Lyon Playfair, recommended that there should be a system in England like the Prussian leaving certificate and that the exam should be 'taken with a quiet mind and without painful effort' (Sessional Proceedings of the National Association from the Promotion of Social Science, 1871–72).

A book written in 1877 by a Cambridge don, Henry Latham of Trinity Hall, *On the action of examinations considered as a means of selection*, provided the following critique of the value of what was being assessed in exams. He looked at the Civil Service exams and claimed that a high class in the exam did not mean necessarily that here was a candidate of high quality. There were two ways in which this came about. He claimed that the exams penalised important qualities such as originality and independence, and that all an exam could do was test knowledge, not mental powers or sound judgement. Secondly, Latham questioned the judgements of the examiners and claimed that their subjectivity affected the results. In reviewing the use of exams in universities, Latham argued that what was being assessed was becoming less important than the struggle of students to attain distinction and the examiners to pick out the best students. The content of the exams was thus becoming a secondary matter, yet they were dominating the teaching curriculum of the university (Montgomery, 1965).

A further strand of criticism emerged after the Bryce Commission (see below) and in some respects in response to the outcomes of that commission. In a Symposium published in 1889, entitled 'The Sacrifice of Education to Examination', the editor, Auberon Herbert, who was an advanced radical and a Liberal MP, attacked exams '... as a tool of centralisation. They increased the power of those who are in control of them' (quoted in Roach, 1971). The instrument, which initially was to have been an agent for creating change in society, had become a tool of the establishment: 'No remedy for existing evils is to be expected by substituting some of these forms of centralisation for others, but only by allowing the utmost freedom for new wants and new forms of thought to express themselves in new systems to compete with the old.' Herbert, as a free-trade economist, was suspicious of all government control and he broadened his argument for decentralisation into opposition to the setting up of an Education Department.

The Bryce Commission

In 1894 a Royal Commission was set up, under the chairmanship of James Bryce, to enquire into the whole subject of secondary education. One of the topics the commission dealt with was the proliferation of exams, and with schools' continued demand for more standardisation. The Cambridge Syndicate's view was presented to the Commission in a memorandum in June 1894. It referred to the original aim of the Local Examinations to improve the state of education in secondary schools and claimed, 'The high character of the work sent in by the pupils at many of the schools which regularly prepare candidates, and the gradual rise which has on the whole taken place in the difficulty of the examinations, afford a satisfactory evidence of progress. The Syndicate believe that this progress may fairly be attributed in a considerable degree to the local examinations themselves.' (Royal Commission, 1895).

The commission reported in 1895 and accepted the important role that examinations played in the system. It recommended the setting up of a central government board and in 1899 the Board of Education was created. Later, in 1902, The Balfour Education Act provided for the establishment of a system of state secondary schools through Local Education Authorities. However, concerning examinations, Bryce took on what was by then the traditionalist point of view, (Montgomery, 1965), that it was preferable to leave examination boards as they were, and to preserve the links with the universities.

The contribution of independent organisations to examining

The introduction of national examinations in this country owes much to an attitude that lay deep within the Victorian view of society: that people should be free to develop themselves. This led to a *laissez-faire* attitude to government intervention in education, but it also left space for the energetic work and enthusiasm of individual educationalists, and of educational institutions. The consensus at the end of the century was that, rather than concentrate power over education in a single government department, such power should be diffused into strong, but accountable, institutions.

How far this was a reason that the work of the mid-nineteenth century pioneers endured needs to be thought through. The focus of that discussion should not be merely on procedural and administrative matters (i.e. who could run the system most efficiently?), but on the educational value of the work that has been done. Exam boards with their roots in universities or other educational institutions, and later with full teacher participation, could represent to people the underlying purpose and meaning of education. Their certificates could express the candidate's relationship to the educational enterprise. This gave them more than utilitarian value. Candidates were linked to the value-base of the academic disciplines and to the communities of those who adhered to those disciplines. In the case of vocational exams, the board (e.g. RSA Examinations) gained its credibility by fostering close relationships with the world of industry and commerce, and by linking students with this world. In a similar way to the academic enterprise mentioned above, its assessments gained their meaning and value from the human activity (commerce) which was their focus.

Independence for exam boards was not the same as licence. They could not act only on their own account, and it was because they were independent that they had to build relationships with their centres and candidates. This sense of relationship extended to that between the examiner and the examinee, in which the examiners tried to do a fair job for each individual student. If the transaction between the examiner and examinee had become purely bureaucratic and impersonal, the system would have been in danger of losing its sense of fairness for the candidate. With the huge increase in examination taking which we now see, this is one of the strong arguments for including a role for teacher assessment in the process, as well as for the need to maintain trust in 'the examiner'.

9 Henry Latham, in his comprehensive book, *On the action of examinations considered as a means of selection*, 1877, deals with this point. He claims that he had found that students who had undergone too many exams 'were, usually, for long time, incapable of giving their minds steadily to any subject requiring close attention', p.48.

The Locals began in the 1850s partly because they were seen to be a fair way to identify and reward ability. Twenty years later the boards were working out their responses to accusations of not being fair. It is how far the present examination system is seen as being fair to individuals which will make the difference between it being perceived as a liberalising or a reactionary one. Independent examining boards are well-placed to respond to this challenge and they continue to play an important part in maintaining the credibility of that system.

References

Bentham, J. (written and printed 1827, published 1830). *Constitutional code: Vol. 1.* London: R.Heward.

English Journal of Education (1857). **XI**, 226.

Foden, F. (1989). *The Examiner, James Booth and the Origins of Common Examinations*. Leeds: School of Continuing Education, University of Leeds.

Headmistress's Reports to Governors (1871). **II**, 256–256.

Herbert, A. (ed.) (1889). *The sacrifice of education to examination*. London: Williams & Norgate.

Latham, H. (1877). *On the action of examinations considered as a means of selection*. Cambridge: Deighton, Bell.

Mill, J.S. (1859). *On Liberty*. London: John W. Parker & Son.

Montgomery, R.J. (1965). *Examinations: An account of their evolution as administrative devices in England*. London: Longmans.

Roach, J. (1971). *Public Examinations in England: 1850 – 1900*. Cambridge: Cambridge University Press.

Royal Commission on Secondary Education 1894. (1895). **V**., 268. London: HMSO.

Sandford, E.G. (ed.) (1906). *Memoirs of Archbishop Temple by 'Seven Friends'*. **II**, 541, 68–69; 549. London: Macmillan & Co. Ltd. Quoted in Roach (1971).

Sessional Proceedings of the National Association for the Promotion of Social Science, 1871–72. 29.

Stephen, B. (1927). *Emily Davies and Girton College*. 83–4. London: Constable. Quoted in Roach (1971).

The Transactions of the Social Science Association (1872). 271–9

Weise, L. (trans. L. Schmitz, 1877). *German Letters on English Education: Written during an Educational Tour in England in 1876*. London: Collins.

Woodard, N. (1848). *A Plea for the Middle Classes*. (pamphlet).

Woodard, N. (1851). *Public Schools for the Middle Classes*. (pamphlet).

PSYCHOLOGY OF ASSESSMENT

Investigating the judgemental marking process: an overview of our recent research

Dr Irenka Suto, Victoria Crisp and Dr Jackie Greatorex Research Division

Prior to Cambridge Assessment's recent interest in the area, the process of marking GCSE and A-level examination questions had received surprisingly little attention among psychologists and social scientists. Whilst there has been some research into marking processes in other contexts (e.g. English essay marking: Barritt, Stock and Clark, 1986, Pula and Huot, 1993; English as a second language: Cumming, 1990, Vaughan, 1991, Milanovic, Saville and Shuhong, 1996, Lumley, 2002) to our knowledge, only Sanderson (2001) has explored the process in depth, producing a socio-cognitive model of A-level essay marking. To address this dearth in knowledge, members of the Core Research team have conducted several linked projects, considering the marking process from different angles. Key aims have been to provide insights into how examiner training and marking accuracy could be improved, as well as reasoned justifications for how item types might be assigned to different examiner groups in the future.

As with any major research question, the issue of how examiners mark items needs to be explored from many different angles to gain as full and cohesive an answer as possible. In biological research, for example, the nature and effects of an illness are explored at multiple levels: molecular, intra-cellular, cellular, physiological, whole organismal, and even epidemiological and demographic. Similarly, some physics researchers conduct fine-grained analyses of minute particles, while others monitor massive structures in space, both in their attempts to establish how the universe began. Linking together these jigsaw pieces in order to see the bigger picture and gain a real overview of a process or phenomenon can be a difficult but necessary challenge. As with biology and physics, this is an important task for researchers in educational assessment.

To recognise the different approaches to research and analysis that the marking process engenders, it is worth considering the very broad research field in which it primarily lies – that of human judgement and decision-making. There exist a number of well-established approaches to investigation, adopted by researchers working within diverse paradigms, and as with the natural sciences, questions are explored on a number of levels. For example, a key approach has been to ask *what information* people attend to and utilise when making decisions. On perhaps the most 'fine-grained' level of research, cognitive psychologists have identified and scrutinised shifts in visual attention among small pieces of information, such as letters, numbers and words on a page of writing. At another level, other psychological researchers have focused on cognitive heuristics and biases in information processing. At yet another level, the influences and roles of behavioural and social information have been explored by researchers interested in such dynamics, and at yet another level still, the effects of physical information in the environment have been studied. Studies at all of these levels have provided important contributions to our understanding in the research field, although there is the potential for them to be integrated much more.

Another popular approach to understanding judgement and decision-making has been to explore the sequences of mental operations by which people arrive at their choices. This approach has proven particularly popular in clinical and legal settings, and again, it has been adopted at a number of levels. In the Core Research team's work on the marking process, we have combined this approach with the one outlined above:

our projects explore both the information that people attend to when marking items and the sequences of mental operations involved.

At a relatively fine-grained level, in our first project, entitled *Markers' Minds*, we identified the cognitive marking strategies entailed in marking GCSE maths and business studies questions (Suto and Greatorex, *in press,* a). This was done using the think aloud method with experienced examiners (Greatorex and Suto, *in submission*). Working within the 'dual-processing' paradigm (Kahneman and Frederick, 2002), we interpreted two of our five strategies (*matching* and *no response*) as utilising simple 'System 1' or 'intuitive' judgemental processes, two strategies (*scrutinising* and *evaluating*) as utilising more complex 'System 2' or 'reflective' processes, and one strategy (*scanning*) as engaging System 1 and/or System 2. An analysis of strategy usage (Greatorex and Suto, 2005; Suto and Greatorex, *in press,* b) revealed that although there are some differences among individual examiners, the most prominent differences occur between subjects and among questions.

A second closely-linked project (Greatorex and Suto, 2006) entailed re-analysing the verbal protocols of 'expert' and 'subject' markers who had marked GCSE maths and A-level physics papers on screen. The data provided further evidence for the five cognitive strategies, indicating that they are used in other marking contexts.

As part of a subsequent project, entitled *Marking Expertise 1*, we explored the relationship between cognitive marking strategy complexity and marking accuracy (Suto and Nádas, 2007a; Suto and Nádas, *in press*). A new theoretical framework was constructed, conceptualising marking accuracy for a particular question as being determined by both (i) a marker's expertise, and (ii) the demands of the marking task. It was proposed that these two factors are in turn affected by a number of different variables, including the complexity of the marking strategies needed to mark the question, and utilisation knowledge of the marking strategies (i.e. knowing which strategies to apply when).

The question-by-question marking of GCSE maths and physics by

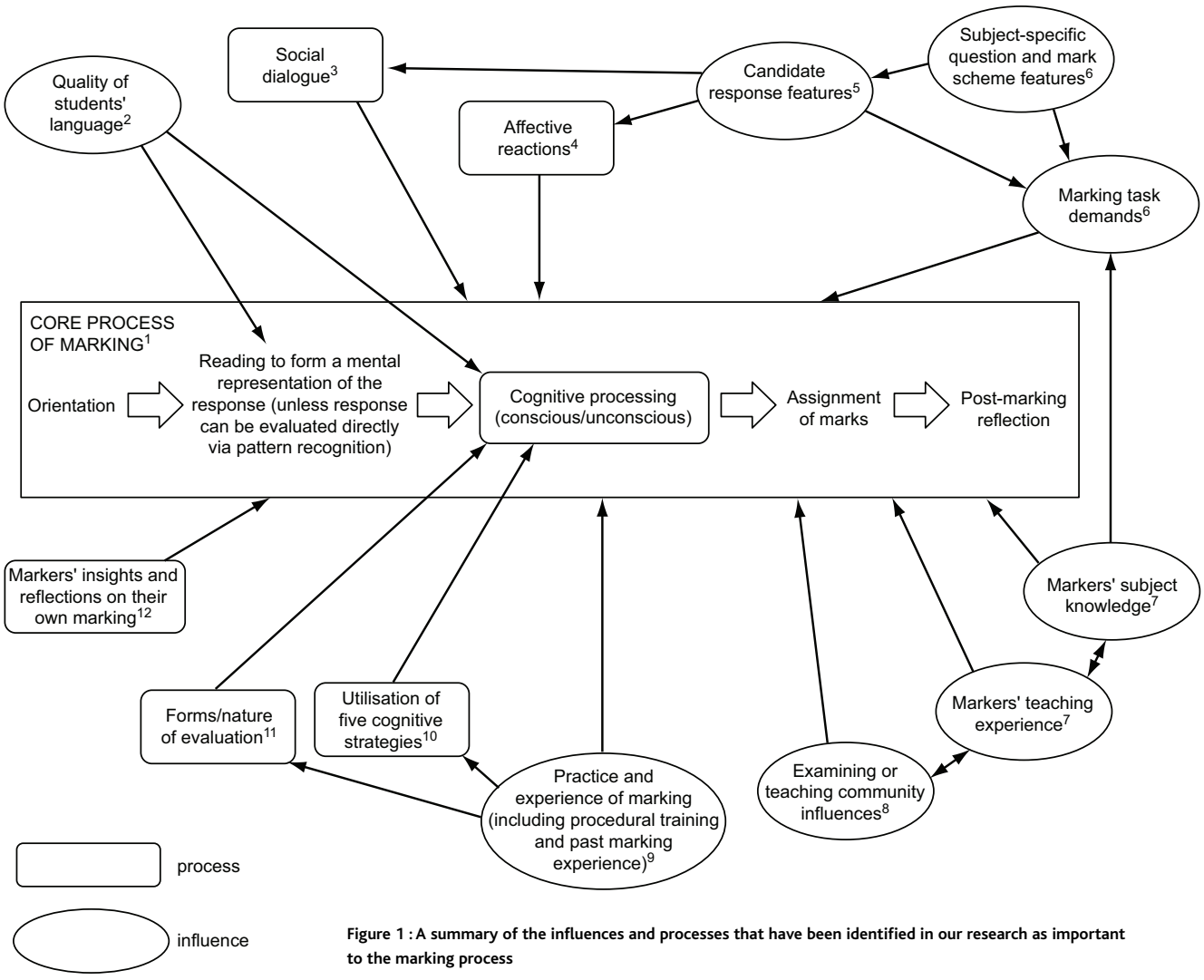


Figure 1 : A summary of the influences and processes that have been identified in our research as important to the marking process

1. Crisp (2007c; *in submission*), Greatorex and Suto (2006), Suto and Greatorex (2006)
2. Crisp (2007a; 2007b; 2007c)
3. Crisp (*in press*)
4. Crisp (2007b; *in press*)
5. Crisp (2007a; 2007b; *in press*; *in submission*)
6. Suto and Nádas (2007a; 2007b)
7. Greatorex, Nádas, Suto and Bell (2007), Suto and Nádas (2007a, *in press*)
8. Crisp (*in submission*), Greatorex, Nádas, Suto and Bell (2007), Suto and Nádas (2007a, *in press*)
9. Greatorex and Bell (*in press*), Greatorex, Nádas, Suto and Bell (2007), Suto and Nádas (2007a, *in press*)
10. Greatorex (2006; 2007), Greatorex and Suto (2006), Suto and Greatorex (*in press*, a, b)
11. Crisp (2007b; 2007c; *in press*; *in submission*)
12. Nádas and Suto (2007), Suto and Nádas (2007a)

'expert' and 'graduate' markers was explored, and it was found that the apparent marking strategy complexity that a question entails is indeed associated with the accuracy with which it is marked. (Instead of using the think aloud method to demonstrate which marking strategies were used, researchers rated the marking strategy complexity of each question a priori.) The finding was replicated in a study of A-level biology marking (*Internalising Mark Schemes*, reported as part of Greatorex, Nádas, Suto and Bell, 2007). Furthermore, one finding was that apparent marking strategy complexity was a useful indicator of how much the standardisation meeting improves marking accuracy; this was the case for two of the three subjects investigated (Greatorex *et al*., 2007). In *Marking Expertise 1*, apparent marking strategy usage was also found to be associated with various subject-specific question features, which are in turn associated with accuracy (Suto and Nádas, 2007b).

These projects have been generally well received, and researchers outside Cambridge Assessment who attended conference presentations of the *Markers' Minds* research have been interested to know how the cognitive strategies relate to other more socio-cognitive perspectives. This question has begun to be addressed in another project: *Holistic versus Structured marking* (Crisp, 2007a; 2007b; *in press*; *in submission*). The primary aim of this research was to compare the process of marking short/medium answer questions with that of marking essays. This was achieved in the context of A-level geography, again using the think aloud method to collect data from examiners. This time, however, the analysis was broader, covering a number of different levels. Several well-established theoretical perspectives were brought into the analysis: constructivist theories of reading comprehension, discourse communities, and communities of practice.

A number of types of examiner behaviours and reactions were identified which were compared between question types within the qualification (Crisp, 2007a; Crisp, *in press*). The framework was also used to explore individual differences among examiners, a considerable number of which were revealed. Possible associations between marker behaviours and lower marker agreement were investigated leading to tentative implications (Crisp, 2007a; Crisp, *in press*). The appropriateness of the features that examiners attended to was also analysed (Crisp, 2007b). A broad socio-cognitive model bringing together the behaviours and reactions observed was proposed to represent the phases (and loops) involved in the process of marking responses. Links between the proposed model and existing psychological theories of judgement were also explored (Crisp, *in submission*).

The programme of research has now investigated marking in GCSE and A-levels in a number of subjects, for a range of question types and from a number of different perspectives. The diagram in Figure 1 summarises the influences and processes that have been identified as important to the marking process from the research conducted so far. The footnotes indicate which papers report on findings in each area.

Whilst the Core Research team have now contributed significantly to an understanding of GCSE and A-level marking, our investigations are ongoing. As part of some doctoral research, the process of marking coursework is being investigated from a socio-cognitive angle. In *Marking Expertise 2* (Suto and Nádas, 2007a), we are continuing to explore the associations between marking accuracy, apparent marking strategy complexity and question features, this time within the context of GCSE business studies and IGCSE biology marking. In other work the judgement processes involved in moderating vocational portfolios (Johnson and Greatorex, 2007) and grading decisions are being explored.

References

Barritt, L., Stock, P. L. & Clark, F. (1986). Researching practice: evaluating student essays, *College Composition and Communication*, **37**, 315–327.

Crisp, V. (2007a). *Comparing the decision-making processes involved in marking between examiners and between different types of examination questions*. A paper presented at the British Educational Research Association Annual Conference, September 2007, Institute of Education, London.

Crisp, V. (2007b). *Do assessors pay attention to appropriate features of student work when making assessment judgements?* A paper presented at the International Association for Educational Assessment Annual Conference, September 2007, Baku, Azerbaijan.

Crisp, V. (2007c). Researching the judgement processes involved in A-level marking. *Research Matters: A Cambridge Assessment Publication*, **4**, 13–18. This article summarises key findings from Crisp (2007a; *in press*; *in submission*).

Crisp, V. (*in press*). Exploring the nature of examiner thinking during the process of examination marking. *Cambridge Journal of Education*.

Crisp, V. (*in submission*). A tentative model of the judgement processes involved in examination marking.

Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing*, **7**, 31–51.

Greatorex, J. (2006). *Do examiners' approaches to marking change between when they first begin marking and when they have marked many scripts?* Paper presented at the annual conference of the British Educational Research Association, 6–9 September, University of Warwick.

Greatorex, J. (2007). Did examiners' marking strategies change as they marked more scripts? *Research Matters: A Cambridge Assessment Publication*, **4**, 6–13. This article summarises some key findings from Greatorex (2006).

Greatorex, J. & Bell, J.F. (*in press*). What makes AS marking reliable? An experiment with some stages from the standardisation process. *Research Papers in Education*.

Greatorex J., Nádas R., Suto W.M.I., & Bell J.F. (2007). *Exploring how the cognitive strategies used to mark examination questions relate to the efficacy of examiner training*. Paper presented at the European Conference on Educational Research, 19–22 September, Ghent, Belgium.

Greatorex, J. & Suto, W.M.I. (2005). *What goes through a marker's mind? Gaining theoretical insights into the A-level and GCSE marking processes*. Paper presented at the 6th annual conference of the Association for Educational Assessment-Europe, November 2005, Dublin, Republic of Ireland.

Greatorex, J. & Suto, W.M.I (2006). *An empirical exploration of human judgement in the marking of school examinations*. Paper presented at the 32nd annual conference of the International Association for Educational Assessment, 21–26 May, Singapore.

Greatorex, J. & Suto, W.M.I. (*in submission*) *What do examiners think of 'thinking aloud'? Interesting findings from an exploratory study*. A version of this paper was presented at the annual conference of the British Educational Research Association, 6–9 September 2006, University of Warwick.

Johnson, M. & Greatorex, J. (2007). *Assessors' holistic judgements about borderline performances: some influencing factors?* Paper presented at the annual conference of the British Educational Research Association, 5–8 September, University of London.

Kahneman, D. & Frederick, S. (2002). Representativeness revisited: attribute substitution in intuitive judgment. In: T. Gilovich., D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: the psychology of intuitive judgment*. Cambridge: Cambridge University Press.

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing*, **19**, 246–276.

Milanovic, M., Saville, N. & Shuhong, S. (1996). A study of the decision making behaviour of composition-markers. In: M. Milanovic & N. Saville (Eds.), *Performance testing, cognition and assessment*. Cambridge: Cambridge University Press.

Nádas, R. & Suto, W.M.I. (2007). *An exploration of self-confidence and insight into marking accuracy among GCSE maths and physics markers*. Paper presented at the annual conferences of the British Educational Research Association, 5–8 September, London; and the International Association for Educational Assessment, 17–21 September, Baku, Azerbaijan.

Pula, J. J. & Huot, B. (1993). A model of background influences on holistic raters. In: M. M. Williamson & B. Huot (Eds.), *Validating holistic scoring for writing assessment: theoretical and empirical foundations*. Cresskill, NJ: Hampton.

Sanderson, P. J. (2001). *Language and differentiation in Examining at A Level*. PhD Thesis. University of Leeds, Leeds.

Suto, W.M.I & Greatorex, J. (2006). A cognitive psychological exploration of the GCSE marking process. *Research Matters: A Cambridge Assessment Publication* **2**, 7–11. This article summarises key findings from Suto, W.M.I & Greatorex, J. (*in press*).

Suto, W.M.I & Greatorex, J. (*in press*, a). What goes through an examiner's mind? Using verbal protocols to gain insights into the GCSE marking process. *British Educational Research Journal*. Findings from this paper were also presented at the annual conference of the British Educational Research Association, September 2005, University of Glamorgan.

Suto, W.M.I & Greatorex, J. (*in press*, b). A quantitative analysis of cognitive strategy usage in the marking of two GCSE examinations. *Assessment in Education: Principles, Policies and Practice*. Findings from this paper were also presented at the annual conference of the Association for Educational Assessment -Europe, November 2005, Dublin, Ireland.

Suto, W.M.I. & Nádas, R. (2007a). The 'Marking Expertise' projects: Empirical investigations of some popular assumptions. *Research Matters: A Cambridge Assessment Publication*, **4**, 2–5. Paper also presented at the annual conference of the International Association for Educational Assessment, 17–21 September, Baku, Azerbaijan.

Suto, W.M.I. & Nádas, R. (2007b). *What makes some GCSE examination questions harder to mark accurately than others? An exploration of question features related to accuracy*. Paper presented at the annual conference of the British Educational Research Association, 5–8 September, London.

Suto, W.M.I. & Nádas, R. (*in press*). What determines GCSE marking accuracy? An exploration of expertise among maths and physics markers. *Research Papers in Education*.

Vaughan, C. (1991). Holistic assessment: What goes on in the rater's mind? In: L.Hamp-Lyons (Ed.), *Assessing Second Language Writing in Academic Contexts*. Norwood, N.J.: Ablex Publishing Corporation.

An exploration of self-confidence and insight into marking accuracy among GCSE maths and physics markers

Rita Nádas and Dr Irenka Suto Research Division

Background

Introduction

A considerable volume of literature in education and occupational research investigates issues in self-confidence and insight, ranging from college students' post-diction self-assessment (e.g. Maki, 1998; Koch, 2001) to work-related self-assessment (Dunning, Heath and Suls, 2004). However, GCSE markers' perceptions of their marking performance and their metacognition have not, to our knowledge, been examined.

Exploring markers' perceptions is important for several reasons. First, if markers' estimates of their own performance prove to be accurate, then this information could be used by Awarding Bodies in standardisation procedures[1] to identify and discuss examination questions that markers have difficulties with. If, however, markers' insight proves to be unreliable and unrelated to their actual marking accuracy, then their feedback on 'problem areas' could be misleading: for example, when conducting standardisation procedures, Principal Examiners might find themselves focussing on the 'wrong' questions. Secondly, investigating whether self-confidence and insight change or become more accurate with more marking practice or more feedback could inform the marker training practices of Awarding Bodies. This may thereby enhance marking accuracy: there is evidence that improvement of one's self-assessment or insight into performance results in enhanced test performance (Koch, 2001; Dunning, Johnson, Ehrlinger and Kruger, 2003).

In this article we present the aims and findings of research which explored GCSE markers' perception of their own marking performance, namely, marking accuracy. Markers' levels of self-confidence and insight and possible changes in these measures over the course of the marking process were investigated. The term 'self-confidence' here denotes markers' post-marking estimates of how accurately they thought they had marked a sample of questions; 'insight' refers to the relationship between markers' actual marking accuracy and estimated accuracy, indicating how precise their estimates were.

Theories of insight and self-confidence

Insight into performance has been widely researched from various angles; and it has generally been found that people tend to have incorrect estimations of their own performance. For example, Dunning *et al*. (2003) found that when asked to predict their mastery on an examination, students in the bottom quartile greatly overestimated their actual performance. They also found that the better performing students were able to predict their raw scores with more accuracy, with top performers actually slightly underestimating their scores.

Several theories have been proposed to explain the phenomenon of poor insight. The nature of self-confidence has been examined by cognitive psychologists, who have adopted the 'self-serving bias' theory. Researchers have found that biases are used by participants in research

1 For regulations on standardisation procedures, see Qualifications and Curriculum Authority, 2006

situations in order to enhance or maintain positive self-views; for example, the *above average* effect (Dunning, Meyerowitz and Holzberg, 2002), or the *optimistic bias / unrealistic optimism* effect (for example, Armor and Taylor, 2002) have been described. Generally, it was found that people tend to have 'overinflated views of their skills that cannot be justified by their objective performance' (Dunning *et al*., 2003).

In some studies, participants were asked to estimate the probability of positive or negative life events that might happen to them (Weinstein, 1980); or to predict their own performance in an imagined or future situation, or *before* completing a task (for example, Griffin and Tversky, 2002). However, participants' actual performances were often not observed in these studies, or feedback was not provided. Thus, studies on self-serving self-assessments have not explored *change* in one's self-confidence after receiving feedback on actual performance. In the few studies in which participants' estimates were compared with their actual performances, results were mixed: while some found that performance estimates and actual performance did not correlate significantly (Griffin and Tversky, 2002), significant, positive and substantial correlations were found by others (e.g., when subjects made correct time estimates for a given task in the study of Buehler *et al*., 1994).

The self-serving bias theory alone cannot explain all findings. It does account for why poor performers tend to give an aggrandised estimation of their own achievement, but fails to reveal why those of higher abilities tend to overestimate their accomplishment to a lesser extent, or why the phenomenon is completely missing in the case of top performers.

The level of someone's self-confidence in their judgements also depends on their social circumstances. Social psychologists (e.g., Sherif, Sherif and Nebergall, 1965) have shown that lay people tend to change their judgements about an ambiguous stimulus when paired with someone who is thought to be an expert in the field, or who seems to be very confident in their judgements: lay people's judgements move in the direction of the expert's judgements. Therefore, the expert is negatively influencing their perceptions of the accuracy of their original judgements, and thus their self-confidence in those judgements. Arguably, the judgements entailed in marking a script could involve a lot of ambiguity for a novice marker: such judgements, and a novice marker's self-confidence in those judgements, are therefore vulnerable to the influences of expert markers' comments. Social influences on markers have been investigated in awarding meetings, where candidates' grades are determined by a team of markers using available script evidence (Murphy *et al*., 1995).

Research into metacognition may also explain why poor insight arises. Metacognition has been widely researched since John Flavell first wrote about it in the 1970s (Flavell, 1979). Cognitive skills are seen to be used to solve a problem or task, whereas metacognition is needed to understand *how* a task was solved (Schraw, 1998). A review of the literature reveals that researchers disagree on the nature of the relationship between metacognition and general cognition; some argue that the same cognitive processes are in the background of both problem solving (for example, marking a script) and also of assessing one's own performance in the given task (Davidson and Sternberg, 1998). This would explain why people with lower cognitive abilities tend to overestimate their test performances (Dunning *et al*., 2003). Others (Borkowski, 2000) describe metacognition as a qualitatively distinct executive process which directs other cognitive processes.

Schraw's theory of metacognition (Schraw, 1998) provides a framework which yields alternative explanations for the findings described earlier, and also a background against which markers' experiences, the marking process, providing self-assessment and receiving feedback can all be comfortably placed. Arguably it is the most comprehensive, therefore, our hypotheses and discussion will be based mainly on this theory. According to Schraw (1998), metacognition is said to have two components: *knowledge of cognition* and *regulation of cognition*. Knowledge of cognition includes three different types of metacognitive awareness: declarative awareness, i.e. knowing *about* things; procedural awareness, i.e. knowing *how*; and conditional awareness, i.e. knowing *when*. Regulation of cognition consists of planning, monitoring and evaluation (Schraw, 1998). These are also the features of metacognition that might differentiate between experts and non-experts in any field.

Arguably, experienced (e.g. 'expert') and inexperienced ('graduate') markers are very different in metacognitive terms. Experts should have extensive declarative awareness (subject knowledge) as they have relevant degrees and normally teach the subjects that they mark. Research suggests they use different cognitive marking strategies for different types of candidate responses (Greatorex and Suto, 2005; Suto and Greatorex, *in press*), therefore, expert markers should have procedural knowledge with extensive conditional knowledge as well. Inexperienced graduate markers, by definition, must also have appropriate declarative awareness (subject knowledge). However, they may lack sufficient procedural knowledge (for lack of opportunity to develop and use efficient marking strategies, for example) and therefore are likely to lack conditional metacognitive awareness as well. Apart from their disadvantage in their lack of knowledge of cognition, inexperienced markers may also lack practice in the regulation of cognition, simply because they have never been involved in the planning, monitoring and evaluation features of the marking process. Therefore, inexperienced markers are likely to have considerably weaker metacognitive skills overall, and it could therefore be expected that they will show less insight into their marking.

However, just like any other cognitive skill, metacognition can be enhanced, among other things, by practice, and this in turn can improve performance (in this case, marking accuracy) (Koch, 2001; Dunning *et al*., 2003).

The 'Marking Expertise' research project

The research explained in this article was originally embedded in a major project on marking expertise (Suto and Nádas, 2007a, b, *in press*). The project examined how expertise and various other factors influence the accuracy of marking previous GCSE papers in maths and physics. The main aim was to investigate possible differences in marking accuracy in two types of markers: experts and graduates. For both subjects, the research involved one Principal Examiner, six experienced ('expert') examiners with both teaching and marking experience and six graduates with extensive subject knowledge but lacking marking and teaching experience. All participants were paid to perform question-by-question marking of the same selections of examination questions collated from previous GCSE papers. The experimental maths paper consisted of 20 questions, the physics paper had 13 questions. Stratified sampling methods were used to select candidate responses for each question, which were photocopied and cleaned of 'live' marks. Two response samples were designed for both subjects; a 15-response 'practice' sample and a 50-response 'main' sample for each question. The marking process for each subject was the following: all markers marked the practice

sample at home, using mark schemes. They then obtained feedback at a single standardisation meeting led by the appropriate Principal Examiner. The main samples were then distributed and were marked from home, and no feedback was given to markers on the last sample.

The marks of the Principal Examiners were taken as 'correct' or 'true' marks and were the basis for data analysis. Three accuracy measures were used: P_0 (the overall proportion of raw agreement between the Principal Examiner and the marker); Mean Actual Difference (MAcD, indicating whether a marker is on average more lenient or more stringent than his or her Principal Examiner); and Mean Absolute Difference (MAbD, an indication of the average magnitude of mark differences between the marker and the Principal Examiner) (for a discussion of accuracy measures, see Bramley, 2007).

Surprisingly, expert and graduate markers were found to be very similar in their marking accuracy both on the practice sample and on the main sample, according to all three accuracy measures. For maths, out of 20 questions in the practice sample, only three showed significant differences between the two types of markers. On the main sample, a significant difference was found on only one question, where graduates were slightly more lenient than the Principal Examiner and experts. For physics, significant differences arose on three questions (out of 13) on the practice sample and on two questions on the main sample. It is worth noting that despite the significant differences, the graduates also produced high levels of accuracy on all questions. There was some improvement in accuracy from the practice sample to the main sample for both groups. As further data analysis showed, the standardisation meeting and marking practice had a beneficial effect on both groups, benefiting graduates more than experts in both subjects.

Aims and hypotheses of the present study

In a further study within our marking expertise research, which is the focus of the present article, we investigated how markers perceived their own marking performance. Our study of insight and self-confidence entailed administering questionnaires at three points during the marking process, and had multiple aims:

Aim 1: To explore experts' and graduates' self-confidence in their marking accuracy before the standardisation meeting.

According to metacognitive theory, and given that graduates are often assumed to be generally less accurate than experts, two hypotheses are plausible; (1) graduates are aware of their lack of metacognitive skills compared with the experts, and they therefore report a lower level of self-confidence after marking the practice sample; and (2) graduates are not aware of their disadvantage, and all participants' self-confidence levels are very similar after marking the practice sample. The first of these hypotheses would seem most probable, as the graduates were informed at the start of the study that expert markers would also be taking part.

Aim 2: To explore changes in experts' and graduates' self-confidence throughout the marking process.

Metacognitive theory would predict that experts' self-confidence would be high throughout the marking process, and might even show a slight improvement, because more marking practice and feedback on the specific exam questions might develop their metacognitive skills as well. It seems reasonable to hypothesise that graduate markers will report rising levels of self-confidence because they should gain marking experience during the process. Therefore, graduates should report increasing self-confidence on each consecutive questionnaire, even to the extent where their self-confidence level reaches that of the experts.

Alternatively, metacognition theory would suggest that graduates' self-confidence levels will drop on the second questionnaire (after the standardisation meeting), for two reasons; first, graduates' judgements might be influenced by the presence of expert examiners at the standardisation meeting, and although they had known about their involvement in the study, expert examiners might have presented a new frame of reference to which to compare their lack of expertise; secondly, they had just received feedback on the Principal Examiner's 'true' or 'correct' marks, and might have had to reconsider their accuracy on the practice sample regardless of the presence of others. This also predicts that graduates' and experts' self-confidence would be the highest on the main sample, and it will be very similar for the two groups.

Aim 3: To explore the initial pattern of insight of experts and graduates, and see whether there are any significant differences between the groups.

Metacognitive theory would predict that only graduates will show poor insight because they lack procedural and conditional metacognitive awareness, while experts should utilise their previous experience in marking and receiving feedback on their accuracy.

Aim 4: To explore whether participants' insight improves through the marking process.

Metacognitive theory would suggest that all participants, but especially graduates should improve their insight with each consecutive questionnaire, because by that time they will have practised marking as well as received feedback (at the standardisation meeting), and will have practised metacognitive skills by giving account of their insight in our questionnaires.

As mentioned earlier, the literature suggests that some researchers see metacognitive abilities as utilising the very same cognitive processes which are used for the problem-solving task itself; others see it as a superior, organising process of other cognitive processes. Since in the first study in our marking expertise project graduates and expert markers were found to be very similar in their performance of marking accuracy (Suto and Nádas, *in press*), we can assume that it is not their basic cognitive abilities which will discriminate between the metacognitive abilities of the two groups (if we find that these differences indeed exist). If this argument is true, then any difference found in the metacognition of the two types of markers could account for differences in the above-mentioned processes (*procedural awareness*, knowing *how*; and *regulation of cognition*, i.e. planning, monitoring and evaluating), rather than for differences in cognitive skills; this could indicate that metacognition and other cognitive processes are not essentially the same phenomena.

Method

Participants

As mentioned previously, 26 markers were recruited: for each subject, six expert markers (with subject knowledge, experience of marking at least one tier of the selected examination paper, and teaching experience), six graduate markers (with subject knowledge but no marking or teaching experience) and one highly experienced Principal Examiner took part in the study.

Procedure

All markers received a letter at the start of the study, informing them that both expert and graduate markers would be participating in the study, and that all markers would mark the same 'practice' and 'main' samples of candidate responses, on a question-by-question basis. Markers filled in questionnaires on three occasions: (1) at the start of the standardisation meeting, after having marked the practice sample (15 responses) at home; (2) after having attended the standardisation meeting; and finally (3) after marking the main sample (50 responses) at home.

In questionnaires 1 (at the start of the standardisation meeting) and 2 (at the end of the standardisation meeting) each marker was asked:

> *How accurately do you feel you have marked the **first** batch [the practice sample] of candidates' responses?*

In questionnaire 3 (after having marked the main sample), each marker was asked:

> *How accurately do you feel you have marked the **second** batch [the main sample] of candidates' responses?*

To each of these questions, the marker had to circle one of the following answers:

1. Very inaccurately
2. Inaccurately
3. No idea
4. Accurately
5. Very accurately

Results

After checking the distributions of the data, mean self-confidence ratings were calculated and t-tests and Mann-Whitney U-tests were used to analyse possible differences between the two types of markers. Pearson's and Spearman's correlation coefficients were calculated to explore whether there were any relationships between actual marking accuracy and the relevant data on self-confidence.

Analysis of self-confidence of expert and graduate markers

Figure 1 shows the mean self-confidence ratings of expert and graduate maths markers on the three occasions when the questionnaires were administered. According to t-tests, graduates and experts differed significantly in their self-confidence ratings of the practice sample in questionnaires 1 ($t = 4.02$, $p < 0.01$) and 2 ($t = 2.87$, $p < 0.05$), where graduates showed significantly lower confidence in their marking accuracy. This difference disappeared in questionnaire 3 ($t = 1.86$, $p > 0.05$); the two marker groups were similar in their estimations of how accurately they had marked the main sample. Change in self-confidence was only found for the graduates, whose self-confidence improved significantly from the first to the third questionnaire ($t = -3.83$, $p < 0.05$).

Figure 2 shows the mean self-confidence ratings of the physics markers. The ratings of experts and graduates were compared. In contrast with maths, no significant differences were identified between the two marker groups on any of the three questionnaires.

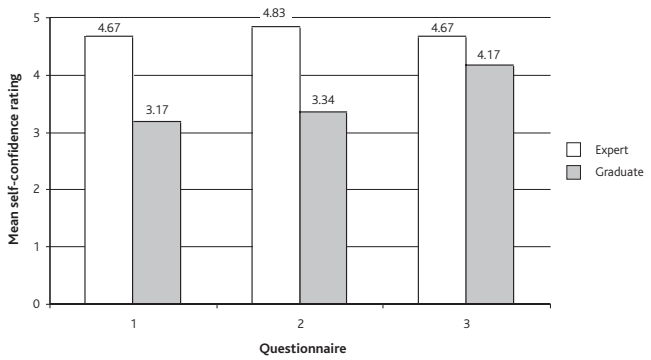


Figure 1 : Graph showing the mean self-confidence ratings of expert and graduate maths markers

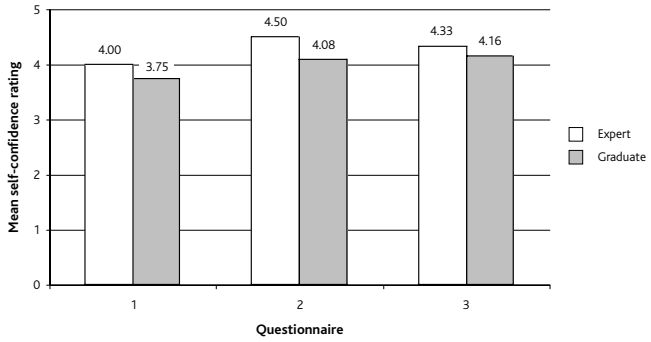


Figure 2 : Graph showing the mean self-confidence ratings of expert and graduate physics markers

Analysis of insight of expert and graduate markers

In order to ascertain whether markers had any insight into their own marking performances, we attempted to correlate the self-confidence data of the two types of markers with their three mean marking accuracy measures (P_0, MAcD, and MAbD) for the practice and main samples.

For maths, Pearson correlation coefficients revealed that neither expert nor graduate markers had real insight into their marking accuracy on either sample; their self-confidence ratings were not significantly related to any of their accuracy measures. The coefficients were the following: for experts: $r = -0.46$, $p = 0.36$ on questionnaire 1; $r = -0.29$, $p = 0.58$ on questionnaire 2; and $r = -0.47$, $p = 0.34$ on questionnaire 3; for graduates: $r = 0.43$, $p = 0.40$ on questionnaire 1; $r = 0.02$, $p = 0.97$ on questionnaire 2; and $r = 0.46$, $p = 0.35$ on questionnaire 3.

For physics, Spearman's and Pearson's correlation coefficients indicated some significant correlations. A significant positive correlation was found for experts' self-confidence after marking the main sample (questionnaire 3) and their mean P_0 values on the main sample ($r = 0.83$, $p < 0.05$) and there was a strong negative correlation with their mean MAbD ($r = -0.86$ $p < 0.05$). Conversely, graduates' self-confidence was significantly negatively correlated to their mean P_0 values ($r = -0.81$, $p < 0.05$) and was positively correlated to mean MAbD values ($r = 0.86$, $p < 0.05$) after the standardisation meeting (on questionnaire 2). Both these correlations indicate that the more accurately the experts marked the main sample, the higher level of self-confidence they reported. Thus, they displayed insight into their own marking accuracies on the main sample. However, the opposite is the case with graduates on the practice sample: the higher self-confidence ratings they gave, the more inaccurate (on two measures) they proved to be. Table 1 summarises the findings.

Table 1: Summary of findings on the correlations between self-confidence levels and marking accuracy

	Does self-confidence on questionnaire 1 correlate significantly with accuracy on the practice sample?	*Does self-confidence on questionnaire 2 correlate significantly with accuracy on the practice sample?*	*Does self-confidence on questionnaire 3 correlate significantly with accuracy on the main sample?*
Maths experts	No	No	No
Maths graduates	No	No	No
Physics experts	No	No	Positive correlation
Physics graduates	No	Negative correlation	No

Discussion

Overall, our results are mixed: our hypotheses were only partially supported by the data, and we found very different patterns of self-confidence and insight for maths and physics markers.

Our first aim was to explore experts' and graduates' self-confidence before the standardisation meeting. All expert markers showed high levels of initial self-confidence; the maths experts' mean level was slightly higher than that of those of both groups of physics markers. It seems that our two hypotheses, namely, that graduates will either report the same level of self-confidence as experts do, or that they will show less self-confidence than that of the experts on the practice sample, applied to one of the graduate groups each: maths graduates showed significantly lower self-confidence than experts, which might reflect expectations of lacking metacognitive and marking skills. Physics graduates, however, showed no difference in their self-confidence from that of experts; in the metacognitive framework this could mean that they did not attempt to account for their lack of experience. However, when these physics graduates' high levels of accuracy are taken into account, their high levels of self-confidence seem only to reflect the expectation of this performance. Finally, it remains a mystery why maths and physics graduates reported different patterns of confidence on the practice sample.

Our second aim was to explore changes in graduates' and experts' self-confidence during the marking process. Metacognitive theory can account for the finding that experts' levels of self-confidence were consistently high; however, no rise was found in their levels of self-confidence over the course of the marking process. Although metacognitive theory would have predicted a small rise, the amount of marking entailed in the study may not have been enough to develop metacognitive skills further. Alternatively, the experts' metacognitive skills may already have been at ceiling level at the start of the research.

As hypothesised, maths graduates were found to report improving levels of self-confidence, up to the point where the significant difference between experts and graduates that had been found previously on the first and second questionnaires disappeared after the main sample had been marked. However, physics graduates were just as confident as experienced examiners were throughout the marking. This is surprising given that graduates, when estimating their own performance, should have taken into consideration their lack of previous marking experience (which they seem to have failed to do on the practice sample already). Nevertheless, they were almost as accurate as experts were, so arguably the equal level of confidence is appropriate but unexpected, as is their high level of marking accuracy.

The data did not support our further hypothesis; the graduates' self-confidence level did not drop after the standardisation meeting in either subject. It seems that the new social reference (expected to be brought about by the presence of experts) or the feedback process did not influence graduates' self-confidence in either subject. However, we did find that all graduates' self-confidence reached the highest level after having marked the main sample, when all previous differences from the experts (if any) diminished.

The third aim was to explore participants' initial insight into their marking accuracy, as indicated by potential correlations between self-confidence and accuracy. Surprisingly, no markers showed any insight on the practice sample before getting feedback at the standardisation meeting. This is especially interesting in the case of expert markers, because metacognitive theory predicts the contrary, counting on their previous experience in evaluating their own marking accuracy. It seems that previous experience in marking different exam questions and in reflecting on one's marking might not generalise to marking new items and to evaluating recent marking accuracy.

Lastly, we explored possible changes in insight in the four marker groups over the course of the marking process. Metacognitive theory would predict that all groups, but especially graduates of both subjects, would improve their insights with each consecutive questionnaire. For maths, surprisingly, neither group showed an improvement in their metacognitive performance with more practice, as neither showed insight on either the practice sample after the meeting, or on the main sample. Data from maths markers, therefore, do not support the metacognitive hypothesis.

For physics, our predictions were, again, only partially supported: experienced markers did show some insight into their marking but only on the main sample. In this case, it seems, the argument that metacognition can be improved by practice was supported by data. Surprisingly, a significant negative correlation was found between physics graduates' estimates and their performance on the practice sample; this, however, seems to support the self-serving bias theory, which predicted this exaggerated optimism. However, the theory predicted the same for all groups, which was not supported by our data.

It has to be noted that because marking was remarkably accurate on the main sample for both experienced and graduate physics markers, we cannot conclude that the difference between their metacognitive abilities is due to different cognitive abilities. Indeed, it may well be that it is the lack of regulation of cognition and procedural knowledge that accounts for different abilities in metacognition. This also sheds light on the nature of the relationship between cognition and metacognition; as graduate physics markers performed similarly to experts on a cognitively

demanding task, but they showed a different pattern of metacognition, this suggests that the two processes might not be essentially the very same phenomena. Of course, further empirical research is needed to examine this point in detail.

Limitations

Just as with all research, our study had some limitations. One of the most obvious ones is that the study involved small groups of participants, which did not allow for the detailed analysis of possible age and gender differences in self-confidence and insight. Participants differed from one another on multiple variables; expert markers had both teaching and marking experience, whereas graduate markers were all young professionals. Also, many of the graduates had attended the University of Cambridge, which might have an effect of its own; for example, Cambridge graduates might be more academically focussed; or more or less conscientious or self-assured than graduates from other institutions. A wider variety of expertise and backgrounds of markers is needed for further research.

A further limitation is that the study involved just two examination papers, which were similar in nature. Using other subjects might have produced different outcomes. Another cause for concern is that there is no way of knowing how seriously markers took our questionnaires; whether they took the time and thought about their confidence in their accuracy overall, or whether they just entered a figure without much self-reflection. This uncertainty also stems from the use of an 'experimental' examination process, created for research purposes only, and the marks given had no effect on any candidate's life chances. Had it been 'live' marking, we might have found different levels of self-confidence and insight. And finally, another source of limitation is that marking practice and metacognitive tasks were always performed at the same time, thus the design of the study did not allow for a separate evaluation of effects; a further study would need the separation of these tasks.

Conclusions and further research

Markers of different subjects show very different patterns of self-confidence and insight. Graduate maths markers showed significantly lower self-confidence than maths experts on the practice sample, but not on the main sample. Physics graduates were as confident as expert markers were throughout the marking process. Generally, markers reported constant levels of self-confidence throughout the marking process; only maths graduates improved their self-confidence from the initial marking of the practice sample to the main sample.

Some markers showed some insight into their marking, but this was not consistent, and even experts' insight was not always accurate. Maths markers showed no insight into their accuracies on either the practice or the main sample. Physics experts showed correct insight on the main sample; graduates showed a significant negative correlation between their performance estimates and their actual marking accuracy on the practice sample.

Because of the mixed results, no one theory fully explains all our data; however, it seems that most, but not all of our results can be interpreted in the framework of the theory of metacognition. Thus, this study also serves as an empirical investigation into the nature of the relationship between cognition and metacognition. Differences in insight between experienced and graduate physics markers did not reflect their overall similarity in accuracy; therefore, differences in metacognitive abilities should reflect differences in procedural and conditional awareness, not cognitive abilities. This suggests that cognition and metacognition may entail qualitatively different processes. It is unclear why maths and physics markers showed such different patterns of self-confidence and insight.

As mentioned in the introduction, one practical implication of this study is for standardisation meetings, where the Principal Examiners and their teams discuss questions on which examiners think they were inaccurate. However, the present study has shown that, especially for maths markers, examiners do not have insight into their own accuracy, therefore they cannot tell which questions should be discussed at the meeting. This could be resolved by on-screen marking, where standardisation procedures can entail immediate feedback on marking accuracy, thereby improving markers' insight; or by conducting qualitative studies (using the Kelly's Repertory Grid technique, for example) which invite Principal Examiners as participants to generate further information on what features of a question make it more difficult to mark than others (see Suto and Nádas, 2007b).

Inquiry into markers' metacognition has been extended in an ongoing follow-up study, where several of the limitations of the first study have been eliminated by a more sophisticated research design. In this experimental marking study, we are looking at how over eighty participants with different background experiences mark business studies GCSE and biology International GCSE (IGCSE) examination papers. Markers' metacognition and aspects of their personalities are being investigated using extended questionnaires. The data analysis of this study is currently under way. We are planning to share our results in 2008.

References

Armor, D.A. & Taylor, S.E. (2002). When predictions fail: The dilemma of unrealistic optimism. In: T. Gilovich, D. Griffin & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgement*. Cambridge: Cambridge University Press, 334–347.

Borkowski, J.G. (2000). *The assessment of executive functioning*. Paper delivered at the annual convention of the American Educational Research Association, New Orleans, April 2000.

Bramley, T. (2007). Quantifying marker agreement: terminology, statistics and issues. *Research Matters: A Cambridge Assessment Publication*, **4**, 22–28.

Buehler, R., Griffin, D., & Ross, M. (1994). Exploring the 'planning fallacy': Why people underestimate their task completion times. *Journal of Personality and Social Psychology*, **67**, 366–381.

Davidson, J.E. & Sternberg, R. J. (1998). How metacognition helps. In: D.J. Hacker, J. Dunlosky & A.C. Graesser (Eds.), *Metacognition in educational theory and practice*. London: Lawrence Erlbaum Associates, 47–68.

Dunning, D., Meyerowitz, J.A., & Holzberg, A.D. (2002). Ambiguity and self-evaluation: The role of idiosyncratic trait definitions in self-serving assessments and ability. In: T. Gilovich, D. Griffin & D. Kahneman (Eds.) *Heuristics and biases: The psychology of intuitive judgement*. Cambridge: Cambridge University Press, 324–333.

Dunning, D., Johnson, K., Ehrlinger, J., & Kruger, J. (2003). Why people fail to recognise their own incompetence. *Current Directions in Psychological Science*, 83–87.

Dunning, D., Heath C. & Suls, J.M. (2004). Flawed self-assessment: Implications for health, education, and the workplace. *Psychological Science in the Public Interest*, **5**, 3, 69–106.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist*, **34**, 906–911.

Greatorex, J. & Suto, W.M.I. (2005). *What goes through a marker's mind? Gaining theoretical insights into the A-level and GCSE marking process*. A report of a discussion group at Association for Educational Assessment – Europe, Dublin, November 2005.

Griffin, D. & Tversky, A. (2002). The weighing of evidence and the determinants of confidence. In: T. Gilovich, D. Griffin. & D. Kahneman (Eds.) *Heuristics and biases: The psychology of intuitive judgement*. Cambridge: Cambridge University Press, 230–249.

Koch, Adina (2001). Training in metacognition and comprehension of physics texts. *Science Education*, **85**, 6, 758–768.

Maki, R. H. (1998). Test predictions over text material. In: D.J. Hacker, J. Dunlosky & A.C. Graesser (Eds.) *Metacognition in educational theory and practice*. London: Lawrence Erlbaum Associates, 117–144.

Murphy, R., Burke P., Cotton, T. *et al*. (1995). *The dynamics of GCSE awarding. Report of a project conducted for the School Curriculum and Assessment Authority*. Nottingham: School of Education, University of Nottingham.

Qualifications and Curriculum Authority (2006). *GCSE, GCE, VCE, GNVQ and AEA Code of Practice 2005/6*. London: Qualifications and Curriculum Authority.

Schraw, G. (1998). Promoting General Metacognitive Awareness. *Instructional Science*, **26** 113–25.

Sherif, C., Sherif, M. & Nebergall, R. (1965). *Attitude and attitude change: The social judgement-involvement approach*. Philadelphia: Saunders.

Suto, W.M.I. & Greatorex, J. (*in press*). A quantitative analysis of cognitive strategy usage in the marking of two GCSE examinations. *Assessment in Education: Principles, Policies and Practice*.

Suto, W.M.I. & Nádas, R. (2007a). The 'Marking Expertise' projects: Empirical investigations of some popular assumptions. *Research Matters: A Cambridge Assessment Publication*, **4**, 2–5.

Suto, W.M.I. & Nádas, R. (2007b). *What makes some GCSE examination questions harder to mark than others? An exploration of question features related to marking accuracy*. A paper presented at the British Educational Research Association Annual Conference, London, 2007.

Suto, W.M.I. & Nádas, R. (*in press*). What determines GCSE marking accuracy? An exploration of expertise among maths and physics markers. *Research Papers in Education*.

Weinstein, N.D (1980). Unrealistic optimism about future life events. *Journal of Personality and Social Psychology*, **39**, 806–820.

The influence of performance data on awarders' estimates in Angoff awarding meetings

Nadežda Novaković Research Division

Background

A variety of standard-setting methods are used in criterion-referenced assessment[1] to decide upon pass scores which separate competent from not yet competent examinees. During the past few decades, these methods have come under close scrutiny not only from the research and academic community, but also from a wider community of stakeholders who have a vested interest in assuring that these methods are the most accurate and fair means of determining performance standards.

The Angoff method (Angoff, 1971) is one of the most widely used procedures for computing cut scores in both the vocational and general education settings. In the Angoff standard setting procedure, a panel of judges with subject expertise are asked to individually estimate, for each test item, the percentage of *minimally competent* or *borderline* candidates (MCCs)[2] who would be able to answer that item correctly.

Within the context of some OCR multiple-choice vocational examinations, judges have the opportunity to make two rounds of estimates. The awarders make the initial estimates individually, at home. Later on, they attend an awarding meeting, at which they take part in a discussion about the perceived difficulty of test items. Furthermore, the awarders receive performance data in the form of item facility values, which represent the percentage of all candidates who answered each test item correctly. Both discussion and performance data are supposed to increase the reliability of the procedure and help judges make more accurate estimates about the performance of MCCs (Plake and Impara, 2001).

After discussion and presentation of performance data, the awarders make their final estimates as to what percentage of MCCs would answer each test item correctly. These percentages are summed across items, and the result is an individual judge's pass score for the test paper in question. The average of individual judges' scores represents the recommended pass mark for the test.

The Angoff method is popular because it is flexible, easy to implement and explain to judges and stakeholders, and it uses simple statistics that are easy to calculate and understand (Berk, 1986; Goodwin, 1999; Ricker, 2006).

However, the validity and reliability of the Angoff procedure have been questioned in recent literature. The main criticism is directed against the high cognitive load of the task facing the awarders, who need to form a mental representation of a hypothetical group of MCCs, maintain this image throughout the entire standard setting activity, and estimate as accurately as possible how a group of such candidates would perform on

1 In criterion-referenced assessment, a candidate's performance is judged against an externally set standard.

2 A minimally competent or a borderline candidate is a candidate with sufficient skills to only just achieve a pass.

a test (Berk, 1996; Boursicot and Roberts, 2006; Glass, 1978; Impara and Plake, 1997; Plake and Impara, 2001).

Some of the criticism has also been directed against the potential undesirable effects of discussion and performance data. During the discussion, awarders may feel pressure to conform to the opinion of the group (Fitzpatrick, 1984, cited in Busch and Jaeger, 1990), while performance data from a small unrepresentative sample of candidates may introduce flaws into the procedure (Ricker, 2006). Furthermore, performance data refer to the entire candidature for the given qualification, while judges are asked to estimate the performance of *minimally competent* rather than *all* candidates. Additionally, some researchers have warned that reliability may be artificially introduced by performance data or discussion by eliminating the variability of individual judgements, whereby the resulting standard may 'no longer reflect judges' true perceptions about the examinee performance' (McGinty, 2005; Ricker, 2006).

Aim

The aim of the study was to investigate the relative effect of discussion and performance data on: (1) the awarders' expectations on how MCCs might perform on a test, (2) the magnitude of change in the awarders' estimates between sessions and (3) the awarders' rank-ordering of items in terms of their relative difficulty.

Design

A group of seven awarders made item facility estimates for two tests of comparable difficulty. They made the first round of judgements for both tests individually, at home. At a later stage, the awarders attended two awarding meetings, one for each test. The meetings took place on the same day. At the first meeting, the awarders voiced their opinions about the quality of Test 1, after which they discussed the perceived difficulty of each test item in turn. Following the discussion, the awarders made the final round of item facility estimates. The second meeting took place one hour after the first meeting; the awarders took part in a discussion, but they were also given the performance data before making the final round of estimates. The second meeting resembled as closely as possible the usual OCR Angoff awarding meetings for Vocational Qualifications. The fact that the awarders received performance data at only one of the meetings allowed us to tease apart the effect of discussion and performance data on their item facility estimates.

The awarding meetings

The awarding meetings were chaired by an experienced Chairperson, who co-ordinated the procedure and facilitated the discussion in the way it is usually done at the OCR Angoff awarding meetings for Vocational Qualifications.

At the start of the first meeting the Chairperson introduced the Angoff procedure and the concept of a minimally competent candidate. He described an MCC as a student who would pass the test on a good day, but fail on a bad day. He also mentioned various ways which could help awarders conceptualise MCCs, for example, thinking about students they had taught. In other words, the awarders were directly encouraged to make estimates about the performance of candidates familiar to them. This is a usual recommendation at the OCR Angoff awarding meetings, and while it helps reduce the cognitive difficulty of the awarders' task, it may result in an increase in the variability of awarders' judgements. The awarders were also told not to make estimates on whether MCCs *should* or *ought* to know the question, but on whether they *would* get the question right.

The awarders were also asked not to mention during the discussion the exact estimate values they had given to the items, although they could say whether they had given a low or a high estimate. This recommendation was given to help reduce the potential influence of more vocal awarders on the decisions of the rest of the panel.

The awarders first voiced their opinions about the test paper in general and its relative difficulty and quality, after which they discussed each item in turn. After each item was discussed, the awarders had the chance to change their original estimates, although there was no requirement for them to do so.

At the start of the second meeting, the Chairperson explained the statistical data that the awarders would get at the meeting, which included the discrimination and facility indices for each item. The awarders were made aware that the item facility values did not reflect the performance of MCCs, but the performance of the entire group of candidates who took Test 2. The Chairperson emphasised the fact that there was no reason for the panel to make their item facility estimates agree with the actual item facility values, but he did mention that the latter were a good indicator of which question was easier or harder compared to other questions in the test.

After the introductory part, the second meeting followed the same format as the first meeting.

Tests

The tests used in the study were two multiple-choice tests constructed from the items used in Unit 1 of the OCR Certificate in Teaching Exercise and Fitness Level 2 (Unit 1 – Demonstrate Knowledge of Anatomy and Physiology). These items were drawn from an item bank, and their IRT (Rasch) difficulty values had already been established. This had several advantages. First, it allowed the construction of two tests of comparable difficulty. Secondly, the pass mark could be established by statistical means, using the information on how students performed on these items in the past. The pass mark for both tests was set at 18.

Test 1, containing 27 items, was completed by 105 students, and Test 2, containing 28 items, was completed by 117 students from centres offering Teaching Exercise and Fitness qualification. The tests were completed as part of another experimental study (Johnson, *in press*), that is, these were not 'real' tests and student performance data were used only for research purposes. Students completed Test 1 after completing Test 2.

Awarders

The awarding panel consisted of three female and four male awarders. These were all experts in the field of Teaching Exercise and Fitness. Two awarders had no experience with the Angoff procedure, while the remaining five had already taken part in an Angoff awarding meeting.

Minimally competent candidates

In order to measure how the awarders' estimates compared to the actual performance of MCCs, we had to identify this group of candidates from all the candidates who took the tests. Remember that the awarders' estimates are supposed to reflect the percentage of *minimally competent* candidates rather than the percentage of *all* candidates who would answer test items correctly.

MCCs were identified as those candidates whose score fell 1 SEM (standard error of measurement) above and 1 SEM below the pass score[3] established by using the item bank data. This is a method similar to the one used in Goodwin (1996) and Plake and Impara (2001).

The first column of Table 1 shows the pass marks for both tests calculated using item difficulty values obtained from the item bank. The second and third columns show the mean score achieved by all candidates and the group of candidates we identified as *minimally competent* respectively. Figures in brackets represent the percentage of the total possible mark.

Table 1 : The average performance of all candidates and MCCs for Tests 1 and 2

	Pass mark	*All candidates*		*MCCs*	
		Mean mark	*N*	*Mean mark*	*N*
Test 1	18 (67%)	17.60 (65%)	105	17.87 (66%)	38
Test 2	18 (66%)	16.04 (57%)	117	17.57 (63%)	46

On the whole, the performance of all candidates was better on Test 1 than Test 2. Johnson (*in press*) ascribed this to the practice effect, since the candidates completed Test 1 after having completed Test 2. However, it is worth noting that four members of the awarding panel voiced their opinion that Test 2 was harder than the usual tests administered for this qualification.

Key findings

Frequency of changes

The awarders made more changes to their original estimates if presented with statistical information about candidate performance than if they only took part in the discussion about the quality and perceived difficulty of the test items. The average number of changes between two rounds of estimates for Test 1 was 5.14 (ranging from 0 to 10 changes per awarder). For Test 2, however, the average number of changes was 11.29, with individual awarders making between 1 and 22 changes.

Rank-ordering of test items

The Spearman rank-order correlation coefficient was used to compare the awarders' estimates to the actual item facility values for the group of candidates identified as minimally competent. This showed how successful the awarders were in predicting which test items the MCCs would find harder and which ones they would find easier to answer. The correlation between the initial estimates for Test 1 and the actual item facility values was weak and non-significant (0.23), and it became weaker after the awarding meeting, at which the awarders took part only in discussion (0.19) On the other hand, the correlation between the initial estimates for Test 2 and the actual item facility values was significant and moderate (0.60), and it became stronger after the second meeting (0.79), when the awarders were presented with performance data. These findings are similar to the ones in Busch and Jaeger (1990), where correlations between the actual item facilities and mean item recommendations increased from one session to the other, after the awarders were presented with statistical information on students' performance.

Awarders' expectations

Table 2 shows the recommended pass marks, calculated by averaging the individual awarders' mean item facility estimates after each round of estimates for both Tests 1 and 2. The figures in brackets represent the percentage of the total possible mark.

Table 2 : The awarding panel's recommended pass marks for Tests 1 and 2 on two rounds of estimates

	Mean mark (all candidates)	*Mean mark (MCCs)*	*Recommended pass mark (Round 1)*	*Recommended pass mark (Round 2)*
Test 1	17.60 (65%)	17.87 (66%)	21 (77%)	21 (77%)
Test 2	16.04 (57%)	17.57 (63%)	20 (71%)	19 (69%)

Table 2 shows that, on average, the awarders' expectations were higher than the actual performance of the group of candidates we identified as minimally competent, as well as the entire group of students who took the test. This applies to both rounds of estimates.

Figures 1 and 2 show the mean actual difference (MD) between the awarders' estimates and the actual item facility values for the group of MCCs on both rounds, for Tests 1 and 2 respectively. The MDs were calculated by subtracting the observed item facility value from the awarder's estimated value. Positive values indicate that, on average, an awarder has mostly overestimated, while negative values indicate that the awarder has mostly underestimated the performance of MCCs. The graphs confirm that the awarders generally expected MCCs to perform better on both tests than they actually did, as indicated by the positive values of the individual MDs.

In order to see whether there was a statistically significant difference between the individual awarders' estimates on each round, an ANOVA was carried out on the data using the following model: 'Actual difference = round + item + awarder + round*item + awarder*round' (the asterisk sign, *, indicates an interaction between two variables).

The ANOVA results for Test 1 revealed that there was a significant main effect of item ($F(26) = 46.30$, $p < 0.001$), and a significant main effect of awarder ($F(6) = 12.87$, $p < 0.001$). There was no significant main effect of round ($F(1) = 0.12$, $p = 0.73$); the mean difference between two rounds was 0.003, which is a small effect size ($d = 0.06$), suggesting that overall the examiners made similar estimates on the two rounds. Furthermore, the analysis yielded no significant interaction between round and awarder ($F(6) = 0.13$, $p = 0.99$).

3 The Standard Error of Measurement estimates how repeated measures of a person on the same instrument tend to be distributed around their "true" score – the score that they would obtain if a test were completely error-free.

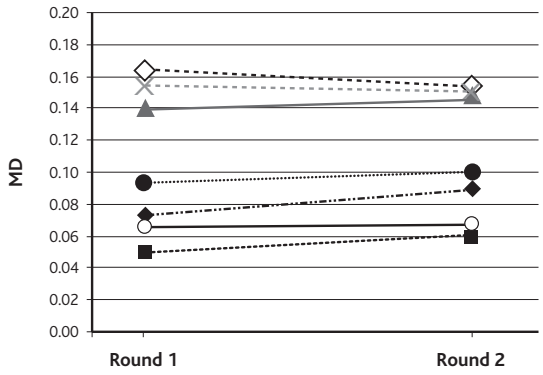


Figure 1: The MD between estimated and actual item facility values on two rounds of estimates for Test 1

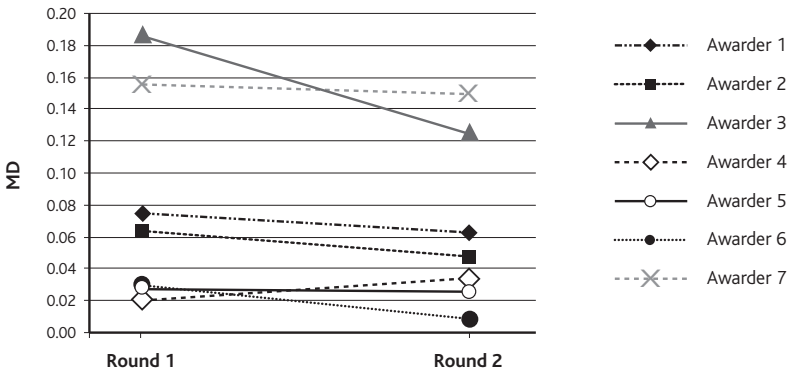


Figure 2: The MD between estimated and actual item facility values on two rounds of estimates for Test 2

On Test 2, the ANOVA revealed a significant main effect of item ($F(27) = 44.75$, $p < 0.001$) and a significant main effect of awarder ($F(6) = 18.79$, $p < 0.001$), indicating that there was a statistically significant difference between individual awarders' MDs. There was no main effect of round ($F(1) = 2.26$, $p = 0.13$); the mean difference between the rounds was 0.015, which is a small effect size ($d = 0.25$). There was no significant interaction between round and awarder ($F(6) = 0.85$, $p = 0.53$).

Figures 3 and 4 show the mean absolute differences (MAD) between the awarders' estimates and the actual item facility value for the group of candidates we identified as minimally competent. Absolute differences were also calculated by subtracting the observed item facility values from the awarder's estimated item facility values. However, all differences were assigned positive values. Absolute differences provide a clear indication of the size of the difference between the awarders' estimates and the actual item facility values.

For Test 1, the results of an ANOVA with MAD as a dependent variable revealed a significant main effect of item ($F(26) = 23.65$, $p < 0.001$) and a significant main effect of awarder ($F(6) = 2.83$, $p = 0.01$). The main effect of round was not significant ($F(1) = 0.04$, $p = 0.84$); the mean difference between rounds was 0.002, which is a small effect size ($d = 0.11$). There was no interaction between round and awarder ($F(6) = 0.03$, $p = 1.00$).

The ANOVA results for Test 2 revealed a significant main effect of item ($F(27) = 17.30$, $p < 0.001$), and a significant main effect of awarder ($F(6) = 2.29$, $p = 0.04$). There was also a significant main effect of round ($F(1) = 7.76$, $p = 0.005$); the mean difference between rounds was 0.026, which is a large effect size ($d = 1.3$). There was no significant interaction between round and awarder ($F(6) = 0.13$, $p = 1$). These results revealed that overall there was a statistically significant change in the size of the MAD between two rounds, although there was no statistically significant difference in the way this changed for different awarders.

Conclusions and implications

The results of the present study support the current OCR practice that awarders at Angoff meetings should be presented with statistical data about candidates' performance.

If the awarders took part only in discussion about the perceived difficulty of test items, the number of changes the awarders made to their initial estimates was relatively small, and there was no change to the pass mark calculated using the initial estimates. Also, there was no statistically significant change from one round to the other, either in the direction or the magnitude of differences between the awarders' estimates and the actual performance of MCCs. Furthermore, the correlation between the awarders' estimates and the actual item facility values for MCCs became weaker after the discussion.

On the other hand, the combination of discussion and performance data had more effect on the awarders' estimates. After being presented with performance data, the awarders made, on average, twice as many changes to their original estimates than when they took part in discussion only. These changes resulted in a statistically significant decrease in the magnitude of differences between the awarders' estimates and the actual item facility values for the group of MCCs.

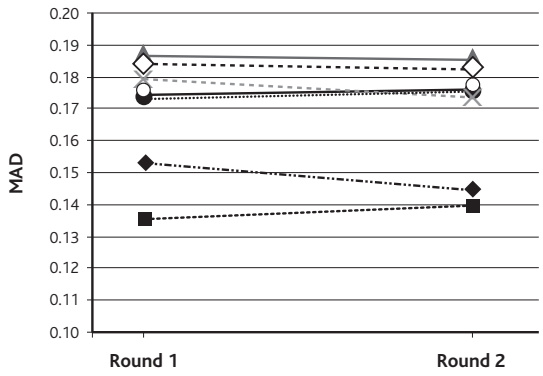


Figure 3 : The MAD between estimated and actual item facility values on two rounds of estimates for Test 1

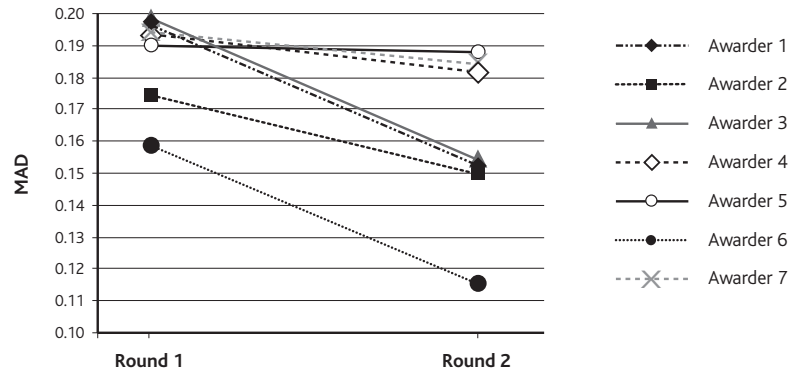


Figure 4 : The MAD between estimated and actual item facility values on two rounds of estimates for Test 2

Furthermore, after being provided with statistical data, the correlation between the awarders' estimates and the actual item facility values became stronger, indicating that the combination of statistical data and discussion helped the awarders judge the relative difficulty of the test items better. However, the provision of performance data had no impact on the direction of differences between the awarders' estimates and the actual item facility estimates.

An important aspect of these findings is that the changes made to the original estimates are observable mostly at the item level. In other words, while the awarders made changes to their item facility estimates, the actual change to the recommended pass mark was rather small (it decreased by only one mark). Furthermore, even after the second round, the recommended pass mark remained three marks higher than the average mark achieved by the total group of candidates. This indicates that the awarders were not swayed in their judgement by statistical data referring to the performance of the total group of candidates who took the test.

Another important finding is that the provision of statistical data does not seem to have affected the variability of awarders' judgements, a concern expressed by some researchers (McGinty, 2005; Ricker, 2006). Generally, if there was a statistically significant difference between the awarders, this difference was observable both before and after the provision of statistical data. In other words, the differences between the awarders were present even after they made changes to their original estimates, indicating that they still maintained their own views about how borderline students would perform on the test, regardless of the actual statistical data they received.

Although the study has provided important and useful findings, there were limitations which must be taken into account when considering its results. The influence of statistical data was tested on only one group of judges who made estimates about test items from a particular examination. However, we do believe that the members of the awarding panel chosen for the study reflect well the experience and expertise of other awarders who take part in the OCR Angoff awarding meetings for various vocational qualifications.

The experimental design of the study was such that only one awarding panel judged both tests, which means there is a risk that the design could be suffering from order effects. Having two awarding panels judging both tests in a different order would be a definite improvement to the present design. Although we had hoped to involve two groups of awarders, we were unfortunately not able to recruit enough participants for this study. Furthermore, the fact that the awarders took part in discussion at both meetings could mean that the discussion they had at the first meeting influenced their judgements at the second meeting as well.

Although the tests used in the study were supposed to be of the same difficulty, the students performed better on one of the tests. Having two groups of students completing the tests in different order would have provided a better indication of whether the better performance on one of the tests was due to the practice effect or whether it could be ascribed to the inherent difficulty of the tests.

It is important to note that the study focused only on some of the aspects of the Angoff method, without attempting to address the broader issues of the validity and reliability of the entire Angoff awarding procedure. These issues could be addressed by rigorous comparison of the Angoff method to other standard setting methods, such as the Bookmark method, for example. Such continuous investigations are necessary to ensure that methods used for setting pass scores are the most reliable, valid, fair and hence the most appropriate to be used both in the context of OCR vocational qualifications, as well as in the context of any standard-based examinations.

References

Angoff, W. (1971). *Scales, norms and equivalent scores*. Washington, DC: American Council on Education.

Berk, R. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. *Review of Educational Research*, **56**, 137–172.

Berk, R. (1996). Standard setting: the next generation (where few psychometricians have gone before!). *Applied Measurement in Education*, **9**, 215–235.

Boursicot, K & Roberts T. (2006). Setting standards in a professional higher education course: Defining the concept of the minimally competent student in performance based assessment at the level of graduation from medical school. *Higher Education Quarterly*, **60**, 74–90.

Busch, J. & Jaeger, R. (1990). Influence of type of judge, normative information, and discussion on standards recommended for the National Teachers Examinations. *Journal of Educational Measurement*, **27**, 2, 145–163.

Fitzpatrick, A. (1984). *Social influences in standard-setting: The effect of group interaction on individuals' judgement*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Glass, G. (1978). Standards and Criteria. *Journal of Educational Measurement*, **15**, 237–61.

Goodwin, L. (1999). Relations between observed item difficulty levels and Angoff minimum passing levels for a group of borderline examinees. *Applied Measurement in Education*, **12**, 13–28.

Impara, J. & Plake, B. (1997). Standard setting: An alternative approach. *Journal of Educational Measurement*, **34**, 4, 353–366.

Jaeger, R. & Busch, J. (1984). *The effects of a Delphi modification of the Angoff-Jaeger standard-setting procedure on standards recommended for the National Teacher Examinations*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Johnson, M (*in press*). Does the anticipation of a grade motivate vocational test takers? *Research in Post Compulsory Education*.

McGinty, D. (2005). Illuminating the "black box" of standard setting; an exploratory qualitative study. *Applied Measurement in Education*, **18**, 3, 269–287.

Plake, B. & Impara, J. (2001). Ability of panelists to estimate item performance for a target group of candidates: an issue in judgmental standard setting. *Educational Assessment*, **7**, 2, 87–97.

Ricker, K. (2006). Setting Cut-Scores: A Critical Review of the Angoff and Modified Angoff Methods. *The Alberta Journal of Educational Research*, **52**, 1, 53–64.

A review of literature regarding the validity of coursework and the rationale for its inclusion in the GCSE

Victoria Crisp Research Division

Introduction

The GCSE was introduced in 1988 and is available in a wide range of subjects. GCSEs are assessed mostly by traditional examination, however, in many subjects a percentage of the assessment is via coursework. In the National Criteria Glossary of Terms coursework is defined as 'all types of activity carried out by candidates during their course of study and assessed for examination purposes' (SEC, 1986, p. 1). In practice, coursework takes a wide range of forms: from written reports of fieldwork in geography, to performances and compositions in music and from pieces of art work, to oral contributions during lessons in English. Coursework tends to involve the assessment of a student's work over a period of time. GCSE coursework (in most cases) is assessed by teachers, internally moderated across teachers within schools and then externally moderated by examiners.

Coursework was included in many GCSEs from their introduction to increase the validity of assessment by providing wider evidence of student work and to enhance pupil learning by valuing skills such as critical thinking and independent learning (SEC, 1985). As the Secondary Examinations Council put it 'above all, the assessment of coursework can correspond much more closely to the scale of values in this wider world, where the individual is judged as much by his or her style of working and ability to cooperate with colleagues as by the eventual product' (SEC, 1985, p. 6). Certain types of subject relevant skills cannot be tested via traditional examinations and the inclusion of a coursework unit as part of the relevant GCSE accommodates the assessment of these skills.

There is continuing debate over whether teachers can be trusted to assess their own students. Some argue that teachers' judgements cannot be free from bias whilst others claim that assessment by teachers is the most valid method (as they see a student's work over a period of time) and that teachers' professional judgements should be trusted. Research evidence shows that the validity and reliability of teacher assessment varies and may be related to certain features such as the degree of specification of tasks and criteria (Harlen, 2004), school cultures (Ellis, 1998) and moderation procedures. Experience suggests that in most cases teachers can successfully rank order candidates' work (although some teachers' marking may be more lenient or more severe than others and require adjustment) and the way that coursework assessment is operationalised and standardised makes use of this fact.

The validity and reliability of the assessment of GCSE coursework has come under much discussion since its introduction with the focus of concerns changing over time. At the inception of the GCSE the main threats anticipated were possible unreliability of teacher marking, possible cheating and concern that girls were favoured (see QCA, 2006a). Now, concerns about consistency across similar subjects, fairness and authenticity (including the issues of internet plagiarism and excessive assistance from others), tasks becoming overly-structured (and hence reducing learning benefits) along with the overall burden on students across subjects, have led to a review of coursework by the Qualifications and Curriculum Authority (QCA). In order to engage with these issues we first need to consider the concepts of validity and reliability.

Validity and reliability

Validity and reliability are central concepts to assessment and describe the confidence we can have in assessment results. Whilst there are slightly different definitions of both reliability and validity, most would agree on the core meanings of these concepts. Reliability is about whether an assessment is repeatable or measures consistently, with a minimum of error. Much attention is given to this issue in assessment development and procedures. The validity of an assessment is about the degree to which it really measures what it purports to measure. Validity and reliability are closely related as a lack of either will result in an assessment that is of little value. In addition, changes to an assessment made to improve validity will often reduce reliability and vice versa.

The traditional view of validity is that there are different kinds of validity: content validity (how appropriate the content of the assessment is as a test of what it aims to assess), construct validity (how well the assessment measures appropriate underlying constructs) and criterion-related validity (how well an assessment relates to actual performance on a specified criterion; this can be predictive or concurrent). In the last few decades most validity theorists have come to consider the construct-content-criterion framework inadequate on the grounds that content and criterion-related validity are actually just examples of evidence that support construct validity. Both Cronbach (1988; 1989) and Messick (1989) consider construct validity the central form. Within this view Messick describes two main threats to construct validity: 'construct under-representation' (the assessment fails to capture important aspects of the construct) and 'construct-irrelevant variance' (capabilities that are irrelevant to the construct are assessed).

Around the same time there was also an emerging view that the concept of validity should be extended to include the consequences of assessment use (Cronbach, 1988; Messick, 1989; Shepard, 1993) specifically with regard to the use of test results, impact on instruction and social consequences. This would include the consideration of whether performance assessment leads to better instructional approaches and does not result in undesirable effects such as narrowing the curriculum (Haertel, 1992). In the climate of both these revisions to the dominant notion of validity, attempts have been made to characterise the types of evidence needed to support construct validity

(e.g. Frederiksen and Collins, 1989; Messick, 1989; Linn, Baker and Dunbar, 1991; Crooks, Kane and Cohen, 1996).

The work of Crooks, Kane and Cohen will be used to provide a structure within which to discuss the validity of GCSE coursework assessment. Crooks, Kane and Cohen's set of criteria has been chosen over those of others as it allows us to focus on the validity that coursework may add as part of a full qualification in comparison to qualifications based only on examinations. In addition, it maps onto key conceptualisations by Messick (1995) and Cronbach (1988) but provides a more practical scaffold for evaluating validity. Crooks, Kane and Cohen (1996) depict assessment validity enquiries as a chain of eight linked stages in order to provide a structure for considering the validity of an assessment. The stages defined are: administration, scoring, aggregation, generalisation, extrapolation, evaluation, decision and impact. For each stage possible threats to validity are exemplified. Crooks, Kane and Cohen suggest that considering possible threats at each stage will allow any 'weak links' to be identified for an assessment.

Validity can be considered a prerequisite to reliability. Crooks, Kane and Cohen (1996) see inter-marker and intra-marker reliability as part of validity because they affect the confidence with which inferences can be made. In the case of coursework, the intention for its use is to improve validity but it may mean greater risks for reliability. Risks to reliability are minimised, at least to some extent, by quality control procedures. However, some teachers initially sympathetic to coursework when the GCSE was introduced were later concerned that the administrative controls put in place to ensure reliability were preventing coursework from being the teacher-led educational experience it should be (Kingdon and Stobart, 1988) and hence limiting the increased validity that coursework was intended to provide.

The validity of GCSE coursework

Although coursework was not a new method of assessment (e.g. it had previously been an optional element of CSEs[1]) it was the introduction of GCSE that saw a much increased presence of coursework as part of the assessment culture through its requirement in most subjects. According to Kingdon and Stobart (1988):

> *...by the time that the GCSE was being introduced, teacher assessment was seen as just another examination technique. Greater understanding of the pros and cons of all techniques had indicated that problems associated with teacher assessment were perhaps no greater than those of other techniques, simply of a different kind. (p. 57)*

The reasons for its introduction were mostly about providing a more valid assessment and allowing the assessment of objectives that cannot be assessed by examination, providing complementary assessment of the same objectives, or to assess objectives for which the evidence is ephemeral (SEC, 1986). As the Secondary Examinations Council state, the aim 'should be one of making what is important measurable rather than of making what is measurable important' (SEC, 1985, p. 2).

Despite the apparent advantages of coursework in terms of validity recent concerns such as the new threat from internet plagiarism led the 2005 *14–19 Education and Skills White Paper* (DfES, 2005) to present QCA with a remit to reconsider the value of coursework and address possible concerns. The initial review (QCA, 2005) involved questionnaires to centres, interviews with teachers, candidates and parents, statistical research and a conference day with examiners. Further work has included a MORI telephone survey of teachers' views (MORI, 2006), a review using this evidence and evidence from QCA monitoring (QCA, 2006a) and an online survey of views on coursework in Maths GCSE (QCA, 2006b).

This section will now use the stages of assessment described by Crooks, Kane and Cohen (1996) to structure discussion of possible improvements to the validity of assessment as a result of including a coursework element within GCSE specifications and possible threats to validity associated with coursework. The stages or links will be considered in reverse order as advised by the authors.

Impact on the student and other participants arising from the assessment processes, interpretations, and decisions

This link in the assessment process as described by Crooks, Kane and Cohen looks at the consequential basis of validity. The direct and indirect impacts of assessment are to be considered along with the effects of experiencing the whole process of assessment. Crooks, Kane and Cohen suggest that threats to validity here can include positive consequences not being achieved or the occurrence of a negative impact of the assessment. The inclusion of coursework in the GCSE was intended to have a positive impact on validity in this respect by providing a number of benefits to learning such as promoting skills of critical thinking, creativity, independent thinking, communication, research and reflection on work (SEC, 1985; SEC, 1986) and allowing helpful feedback from teachers (Ogborn, 1991). Coursework was also intended to be motivating through the realistic sense of audience, the opportunity to negotiate the task and continual assessment (SEC, 1985; SEC, 1986; Ogborn, 1991). In addition, Ogborn (1991) argues that coursework forces teachers to plan courses carefully. In these ways the use of coursework might reduce some threats to validity to do with impact that may exist where assessment consists of examinations alone (e.g. focusing on factual knowledge at the expense of higher order skills). However, if concerns about coursework becoming overly formulaic and predictable in some subjects are well-founded, then coursework may not achieve its intended positive impact. Achieving positive impacts may also be at risk if some students only engage with coursework tasks at a surface level.

Additionally, the heavy workload for teachers and students reported by some constitutes a negative impact of coursework for some of those involved and hence may threaten validity in this respect. In the early days of the GCSE efforts were made to address this concern and various teachers and Local Education Authority professionals investigated and sought to provide advice and good practice ideas based on experience. The main means of controlling the demand of coursework is thought to be to 'ensure that coursework is integrated into the curriculum' (SEC, 1985, p. 8) with tasks arising out of good classroom practice (Cunningham, 1991). Possibilities such as using one piece of coursework to address requirements of more than one subject (Leonard, 1991) or to use methods other than writing were tested but did not become common practice (except for the current overlap in coursework between English and English Literature). Cross-curricular schemes required extra planning from teachers but did reduce student workloads (Leonard, 1991).

It is interesting to note that Scott (1990) found that only a small number of pupils were doing excessive amounts of coursework and other homework. He also reported that the way that pupils reacted to coursework and homework pressure was not related to the amount they actually had to do.

1 CSE was a predecessor of the GCSE.

Decision on actions to be taken in light of judgements

Crooks, Kane and Cohen's 'decision' link is about actions that are taken as a result of judgements, for example, whether a student's score is considered appropriate to admit them to a course. When evaluating the validity of an assessment this stage involves evaluating the merit of the decisions that are taken and whether they are consistent with the information on which they are based and have generally beneficial consequences. One possible threat to validity at this stage according to Crooks, Kane and Cohen would be poor pedagogical decisions. The inclusion of coursework actually gives space for teachers to make good pedagogical decisions. They have more scope to provide useful feedback to students and greater freedom and flexibility within the curriculum, the latter of which was reported by teachers in MORI's survey for QCA (MORI, 2006). However, there is a risk that some teachers may dedicate too much time to coursework at the expense of other areas of study.

Evaluation of the student's performance, forming judgements

This link in the assessment chain is about evaluating what the scores relating to the target domain mean, for example, evaluating what the scores tell us about a student's strengths and weaknesses. Potential threats to validity at this stage can include biased interpretations of test scores (e.g. as a result of a 'halo effect') and poor understanding of the assessment information and its limitations. These issues are the same for GCSE results regardless of whether coursework formed part of the assessment and are hence beyond the scope of the current discussion.

Extrapolation from the assessed domain to a target domain containing all tasks relevant to the proposed interpretation

In the extrapolation link we consider the validity of extrapolating assessment results from the assessed domain to the target domain. This might usually be termed 'generalisability'. According to Crooks, Kane and Cohen, overly constrained assessment conditions would threaten validity in the extrapolation link. This threat to validity is likely to be reduced by the inclusion of a coursework element as part of a qualification.

Another potential threat to validity in terms of extrapolation occurs if parts of the target domain are not assessed or are given minimal weight. This is similar to 'construct under-representation' as described by Messick (1989). The inclusion of coursework in GCSE assessment is likely to improve validity in this respect as it allows types of skills that cannot be assessed by an examination to be evaluated. Improving construct representation was one of the key aims of including coursework in GCSE from the outset.

Avoiding construct under-representation is just as important today as it was when GCSEs were introduced but it seems to be that other threats to validity are currently considered greater concerns and are resulting in changes in the use of coursework.

Generalisation from the particular tasks included in a combined score to the whole domain of similar tasks (the assessed domain)

This link considers the accuracy of generalising from an aggregated score in an assessment to performance in the entire assessed domain (e.g. the entire range of tasks falling within the specification). If the conditions of the assessment vary too much then this can make such generalisations problematic. The term reliability would often be used to describe this issue. With coursework, the conditions do vary somewhat and the tasks used vary but this may be necessary in order for coursework to achieve its purpose of broadening the skills assessed without becoming so over-defined that the learning benefits are lost and risks of plagiarism are increased. The assessment of only a small sample of student work would also threaten reliability. Coursework can involve just one or two tasks but these are large tasks conducted over a longer period of time so they effectively increase the sample size for a GCSE qualification more than could be achieved using an equivalent exam and hence should help to avoid 'construct under-representation' (Messick, 1989).

Aggregation of the scores on individual tasks to produce one or more combined scores (total score of subscale scores)

Issues under Crooks, Kane and Cohen's aggregation link include aggregating tasks that are too diverse and giving inappropriate weights to different aspects of assessment. Whilst the aggregation of scores from coursework and other examined components to determine GCSE grades could be considered an aggregation of diverse tasks, this is not generally considered a problem for the use of coursework. If anything, it is a strength since a wider range of relevant skills can be assessed.

Scoring of the student's performances on the tasks

With regard to the scoring of an assessment, Crooks, Kane and Cohen suggest consideration of aspects that can reduce the validity of score interpretations and consequent decisions. One potential risk to the validity of an assessment in this link is that scoring might fail to capture important qualities of task performance. As Crooks, Kane and Cohen describe 'attempts to increase rater agreement by using more objective scoring criteria will often lead to a narrowing of the factors included in the scoring, thereby increasing the risk posed by this threat to validity' (p. 272). This is something that needs to be kept in mind in the context of the design of coursework guidance and mark schemes in individual GCSE subjects. Coursework assessment offers an improvement on examinations in that there is less risk of scoring emphasising unimportant but easily rated aspects of student performance. However, whilst it has been argued that providing wider evidence of pupil work through coursework will increase the repeatability of the assessment (SEC, 1985; SEC, 1986), it was always acknowledged that monitoring the marking reliability associated with GCSE coursework assessment would be important. Indeed, many of the negative responses to the introduction of GCSE involved fears that coursework marking would be unreliable and easily open to abuse (Kingdon and Stobart, 1988). Leonard (1991) discusses the 'tension between trusting the professional judgement of teachers and the issue of public confidence in the system of assessment' (p. 10). It is perhaps counter-intuitive to public opinion that teachers can judge their own students without bias.

Some data are available on the reliability of coursework marking. Taylor (1992) asked two moderators to re-mark pieces of coursework in each of GCSE English, maths and history and A-Level psychology and compared the marks given between the two moderators with the mark given by the original moderator. Good correlations between different pairs of moderators were found in each subject (ranging from 0.73 to 0.97). Additionally, Taylor found evidence that there were many more centres that over-marked candidates than under-marked. Wiliam (1996) mentions evidence that in the marking of the 100 percent coursework English GCSE teachers learnt to agree on what grade a piece of coursework was worth but they did not always agree on the aspects of the work that were most significant in making the work worth a particular grade.

It is interesting that Crooks, Kane and Cohen comment when discussing potential marker consistency that 'it is desirable to reduce the extent of such inconsistency, but not at the expense of eliminating or reducing the weight given to important aspects of task performance which can only be assessed through professional judgement' (p. 272).

Administration of the assessment tasks to the student

The conditions under which students take an assessment can impact on the validity of interpretations about the assessment and this link in Crooks, Kane and Cohen's model involves examining the task administration. The use of coursework eases the threat to validity caused by stress in exams and is thought to improve motivation. For example, coursework is thought to be fairer for hard-working pupils who are affected by exam stress and also allows the use of tasks that would cause anxiety in an exam situation (SEC, 1985; SEC, 1986). However, the testing conditions involved in coursework can be dissimilar (Scott, 1990) and clashing deadlines for coursework completion across subjects may cause anxiety for some students.

The threat to validity that seems to be considered most significant currently comes under the category of 'administration' and is about ensuring authenticity of student work. As a result of such concerns coursework is currently being reviewed by the Qualifications and Curriculum Authority (QCA). Concerns relate to plagiarism and excessive assistance from others in particular. The arrival of the internet and increased presence of computers in homes has made the potential for plagiarism greater. Additionally, the level of structure and uniformity of coursework tasks may make plagiarism easier.

Some engagement of parents in their child's coursework is encouraged. QCA's first review report (QCA, 2005) found that nearly two-thirds (63%) of parents helped in some way (e.g. checking spelling and grammar, helping to find an article) and 5% of parents with children taking GCSE admitted to actually drafting some of their child's coursework. The report suggests that there is a lack of awareness that this is not allowed and that there are consequential penalties. Such collusion was always a possibility with coursework but seems to be greater concern now than in the past.

The QCA review (2005) reports that some students admitted trying to download assignments from the internet but not to using them. Some admitted having submitted the work of a sibling or friend as their own. There is also a possibility for inadvertent collusion between peers where part of fieldwork or investigations involves group work or identical tasks.

The QCA (2005) report makes a number of proposals including ensuring that teachers can confirm authenticity, guidelines for teachers and parents on limits of permitted help (these have now been prepared and made available) and giving a higher profile to malpractice. These may help to reduce potential threats to validity in this link.

Strong and weak validity links for coursework

Using Crooks, Kane and Cohen's model we can identify the links where coursework reduces threats to validity compared with examinations alone and links where threats to validity remain for coursework. Coursework has strengths in terms of improving construct representation (extrapolation), the potential for positive effects on learning (impact) and increasing motivation and reducing assessment anxiety (administration). When GCSEs began, the threats to validity that caused concern were possible negative effects in terms of impact due to workload for teachers (impact) and the potential for biased or inconsistent marking by teachers (scoring). Recently, concerns have shifted towards the issue of authenticating work (administration) and it is this threat to validity, combined with workload issues for students and teachers that seem to be central in driving current changes.

The future of GCSE coursework

As mentioned earlier, the *14–19 Education and Skills White Paper* pointed to concerns about GCSE coursework and gave QCA the remit of addressing certain issues. The 2005 QCA report concluded that the use of coursework needs review in a number of subjects but that it may not be needed in some subjects. A series of reviews were instigated starting with one focussed on mathematics (given that 66% of teachers felt mathematics coursework was problematic) and a MORI study of teachers' views across seven subjects (QCA, 2006a). The QCA has now confirmed that coursework will be dropped from GCSE mathematics from courses beginning in September 2007 and from a number of other subjects (business studies, classical subjects, economics, English literature, geography, history, modern foreign languages, religious studies and social sciences) from courses beginning in September 2008 where they will be replaced with controlled assessments. Controlled assessments are likely to involve tasks being set or approved by the awarding body, conducted under supervised conditions and marked by teachers (QCA, 2007). This would mean a reduction in possible threats to validity in terms of authentication (administration link) and perhaps in terms of marking reliability (scoring link). However, it could have the potential to reduce the validity benefits of coursework in terms of construct representation (extrapolation link) if tasks limited the skills tested, or to reduce validity benefits in terms of impact if tasks became less interesting or overly structured. It is difficult to be sure of the likely effects on validity until the exact nature of controlled assessments is known.

Decisions over changes have been justified by QCA on the basis of three key principles: that the intended learning outcomes in the subject should be critical in determining the appropriate form of assessment, that the most valid (including reliable) form of assessment for a learning outcome should be used so that results are fair and robust and maintain confidence, and that the assessment process should be manageable (QCA, 2006a). It is interesting that the Heads of Department interviewed by MORI (2006) were fairly positive about coursework, particularly in subjects with oral or practical coursework tasks, and nearly all acknowledged the benefits to students. Furthermore, the QCA reviews report a general consensus of the positive impact of coursework on teaching, learning and assessment and that the benefits outweigh the drawbacks (QCA, 2005).

Concerns about internet plagiarism were not as great as might have been expected (82% of teachers disagreed that students used the internet too much) and whilst more than half felt that students in some schools can gain unfair advantage in the current system the most frequently mentioned drawback was the burden of marking coursework. The interviews by MORI found that 66% of teachers were opposed to removing coursework and 51% were strongly opposed to its removal. The MORI interview evidence would not seem to support the decisions that have been made though the controlled assessment proposals might well address teacher concerns that removing coursework would impact on teaching (e.g. lead to less time spent on practical tasks or fieldwork).

It seems that concerns about threats to validity in the administration

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

the observed score is in agreement with that generated by the candidate's true score (if we knew it). 'Consistency' refers to the proportion of examinees that would be classified the same on another, parallel form of the test (or for our purposes, classified the same by a different marker in the same test). The indices we are interested in are those relating to classification *consistency*, since we do not know the 'correct' mark. In this paper we ignore the impact of other sources of error attributable to the examinee, the particular test questions, etc.

The simplest consistency index is the proportion of candidates (P_0) getting the same grade from the two markers. As an illustration, the following cross tabulation shows the proportion of candidates given each grade by the two different markers, x and y, with an inter-marker correlation of 0.995:

Table 1 : An example cross tabulation of proportions of candidates awarded each grade (simulated data)

y grade	*x grade*						
	A	*B*	*C*	*D*	*E*	*U*	*Total*
A	0.160	0.010	<0.001	0	0	0	0.170
B	0.010	0.088	0.014	<0.001	0	0	0.111
C	<0.001	0.014	0.109	0.016	<0.001	0	0.138
D	0	<0.001	0.016	0.117	0.016	<0.001	0.149
E	0	0	<0.001	0.016	0.152	0.013	0.181
U	0	0	0	0	0.013	0.239	0.252
Total	**0.170**	**0.111**	**0.138**	**0.148**	**0.181**	**0.252**	**1.000**

Hence the proportion of candidates consistently classified is the sum of the diagonal values (P_0=0.865) and therefore the proportion inconsistently classified is 1-0.865 = 0.135.

Please (1971) used this method of measuring misclassification in terms of the difference between the observed grade and the true grade. Thus, he was referring to a measure of classification accuracy and not classification consistency.

He estimated levels of misclassification using this method with reliability coefficients of between 0.75 and 1 for A-levels (on the assumption of a known fixed percent getting each grade – 10% getting A, 15% getting a B etc[1]). For example, with a correlation of 0.93 between true and observed score (and thus reliability, the square of the correlation, equal to 0.865) only 74% of A grades were classified correctly with 24% getting a B and 2% a C. For an exam with reliability of 0.83 or less, more than half the candidates would be wrongly graded. He determined that a reliability of 0.97 was required before less than 25% would be wrongly graded.

Two other UK authors (Cresswell, 1986; Wiliam, 2000) also looked at the reliability of tests by simulating data and reporting the proportion of candidates with the same observed and true grades (although Wiliam actually reported the percentage *incorrectly* classified). By comparing observed score with true score classifications, they were again looking at classification accuracy, not consistency. Both papers showed that increasing the reliability of the test increases the proportion correctly classified, and that increasing the number of grades or levels reduces the proportion. This second conclusion makes intuitive sense, merely because there are a larger number of categories into which to be misclassified.

As Cresswell points out however, increasing the number of grades has the compensatory factor of reducing the severity of any misclassification. For instance, misclassification by one grade on an exam with ten different grades is less serious than a misclassification on an exam with only two grades (pass/fail).

Livingston and Lewis (1995) used the mark distribution on one form of a test to estimate classification consistency on an alternate form. However, they did not look at the overall level of classification, but at the level at each of the grade boundaries in turn. Thus at grade B, the inconsistently classified candidates would be those that would be awarded *at least* a B on one form of the test (marker x in our case), but would get a lower grade from another form (marker y). This gives a series of 2x2 contingency tables for each grade. Using the data from Table 1 we have:

Table 2 : 2x2 contingency tables of proportion of candidates classified at A and B boundaries

x grade	*y grade*	
	A	*B–E*
A	0.160	0.010
B-E	0.010	0.820

inconsistent classification =0.01+0.01= 0.02

x grade	*y grade*	
	A,B	*C,D,E*
A,B	0.268	0.014
C,D,E	0.014	0.705

inconsistent classification =0.014+0.014= 0.028

This index is relevant for UK exams when considering what the results of GCSE or A-level exams are used for: for instance, GCSE results are often summarised in terms of the number getting 5 grade C or above, in which case a candidate misclassified from a grade C to a B or A is less serious than one misclassified from a C to a D. Similarly, A-level results are often used to select candidates for university. The index could then be used to measure candidates who would have been awarded grades good enough to achieve the university's offer by one marker, but not by another.

Lee, Hanson and Brennan (2000) used three different models to simulate data. For each they estimated classification consistency indices, which were calculated for all of the grade boundaries at once or each boundary separately. They also calculated the above indices dependent on the true score. These had the unsurprising outcome that on true scores around where the cut-off points lay the levels of inconsistent classification were higher than on scores in the middle of the categories.

Methodology

We generated a set of exam scores from two hypothetical markers, such that the correlation between the two sets of marks was at a certain level. This was done by simulating a large[2] set of normally distributed data (with mean zero and unit standard deviation): this was the data for marker x. Another set of normally distributed data was generated which correlated with the first set of data to a certain level (say 0.90): this was the data for the second marker (y). Both sets of data were then un-standardised using real means and standard deviations based on past live exam data. This converted the data into two possible sets of normally distributed marks based on the real mean and standard deviation for that

1 In 1971 the number of grades available at A-level was different than today, being A, B, C, D, E, O and F.

2 N>1,000,000. In the scatter plots (Figures 1 and 2) the number of data points has been reduced for clarity.

subject/unit and with the required level of correlation between the two hypothetical markers. This is represented graphically in Figure 1. It should be noted at this point that the simulated data gave both markers the same severity or leniency. The correlation between two examiners, one who marks consistently higher than the other may be very high, but would tend to lead to more inconsistent classification than with two markers with the same level of severity. However, the impact of this is beyond the scope of this article.

The next step was to add in the grade boundaries on both axes. By using the actual boundaries that were used for awarding we determined the number and proportion of candidates that might have been awarded a different grade if their script had been marked by a different marker, for a given level of correlation:

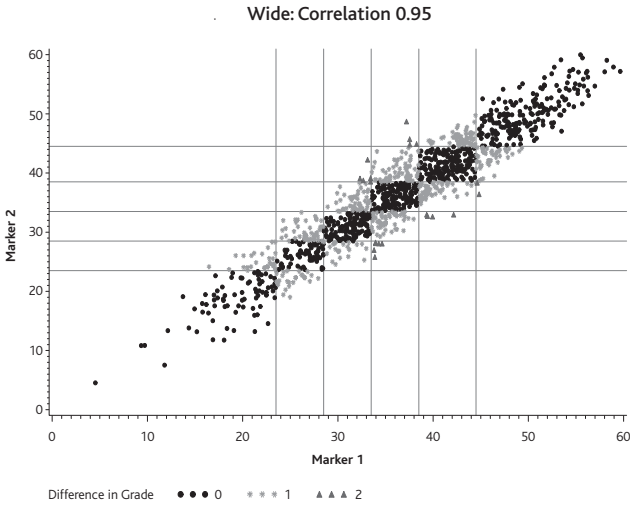


Figure 1 : Scatter plot of marks from two hypothetical markers with grade boundaries (inter-marker correlation = 0.95).

Inspecting the graph gives an idea of the proportion of candidates getting a different grade depending on which marker marked their paper. The candidates who received the same grade are the dots, those who received one grade different are the triangles, and two grades different are stars. The precise proportions of consistent and inconsistent classifications are shown later in Tables 4 and 5.

The next step was to vary the level of inter-marker correlation. It is well documented that this varies between subjects (e.g. Murphy, 1978, 1982; Newton, 1996; Vidal Rodeiro, 2007). Papers with a large number of highly structured questions (Maths, Physics, etc) generate higher correlations than those with a few long answer essay type questions (English, History, etc). This suggests the amount of inconsistent classification will also be different, with a higher level in subjects with lower correlation. Thus we simulated data at different levels of correlation (0.995, 0.95, 0.90, 0.80 and 0.70) and recorded the effect on the amount of inconsistent classification. This is further complicated by the number of grade boundaries and where they lie within the mark range. The closer together the grade boundaries are, and the more grades there are, the more candidates are likely to be inconsistently classified. For example, in an A-level unit with five boundaries all with a width of five marks, the A-E mark range is 25 marks. If a candidate's two scores from the hypothetical examiners differed by three marks then there is a good chance they will get a different grade from each marker, but there is still a fair chance that their classification would be the same under both markers. Now take a unit with grades that are only three marks wide, an A-E mark range of 15 marks. Our candidate with the three mark difference is now sure to get a different grade from each marker. For the same reason, a subject with a narrower grade bandwidth (but the same score distribution) will generate more inconsistent classifications. Whilst it would have been possible to examine the 'pure' effect of changing the grade bandwidth on the same set of simulated data, we felt this would be somewhat unrealistic, since in practice the grade bandwidths depend on the score distributions. Therefore we carried out simulations based on real data for two different subjects, with different A–E bandwidths, and compared the levels of inconsistent classification in each. It is important to emphasise that the 'narrow' and 'wide' units differed in more than the width of their grade bands. Table 3 shows that they also differed in terms of mean, standard deviation and the percentage of candidates in each grade. Therefore comparisons between them in terms of classification consistency do not show the 'pure' effect of spacing of boundaries. However, they do illustrate two realistic scenarios with some interesting contrasts.

Some factors may have a double effect on the inconsistent classification. Increasing the length of an exam for instance is likely to reduce the problem in two ways. First, longer tests tend to increase the inter-marker reliability (Murphy, 1978, 1982) and secondly a longer test is likely to have boundaries that are more separated.

Two A-level units were chosen for this research (from the June 2006 session); both with the same maximum mark but one with relatively closely spaced grade boundaries (A–E width of 13 marks) and one with relatively widely spaced grade boundaries (A–E width of 21 marks). Descriptive data for the two units are shown in Table 3 below.

Table 3 : Descriptive data for the units used

	Narrow		*Wide*	
Candidates	5296		12543	
Max marks	60		60	
Mean	31.86		36.99	
SD	8.01		9.60	
Boundary	*Cut Score*	*% in grade*	*Cut Score*	*% in grade*
A	40	17.75	45	22.72
B	37	11.78	39	23.58
C	34	13.78	34	19.91
D	31	14.41	29	14.68
E	27	16.79	24	10.09
U	0	25.49	0	9.03

We looked at the potential number of candidates inconsistently classified in both units, for different levels of correlation.

Results

We first confirmed that the data we generated could reasonably have come from a real application of the exam by comparing the score distributions generated by each of the simulated markers with the real distribution. Because the simulated data were normally distributed, some observations were above the maximum or below the minimum mark. These were excluded from the analysis. Also, the observations generated were not whole numbers and thus needed to be rounded. These two adjustments had the effect of very slightly altering the mean and standard deviations of the simulated distributions and the correlation

between the two simulated markers. However, these differences were such a small magnitude that they can safely be ignored.

In Table 4 below, P_0 is the overall level of classification consistency, (the sum of the diagonal elements in the cross-tabulations) for the two units at different levels of correlation.

Table 4 : Proportion of candidates consistently classified at different levels of correlation

	Correlation	P_0
Narrow	0.995	0.865
	0.99	0.809
	0.95	0.616
	0.9	0.523
	0.8	0.429
	0.7	0.372
Wide	0.995	0.881
	0.99	0.832
	0.95	0.637
	0.9	0.528
	0.8	0.418
	0.7	0.356

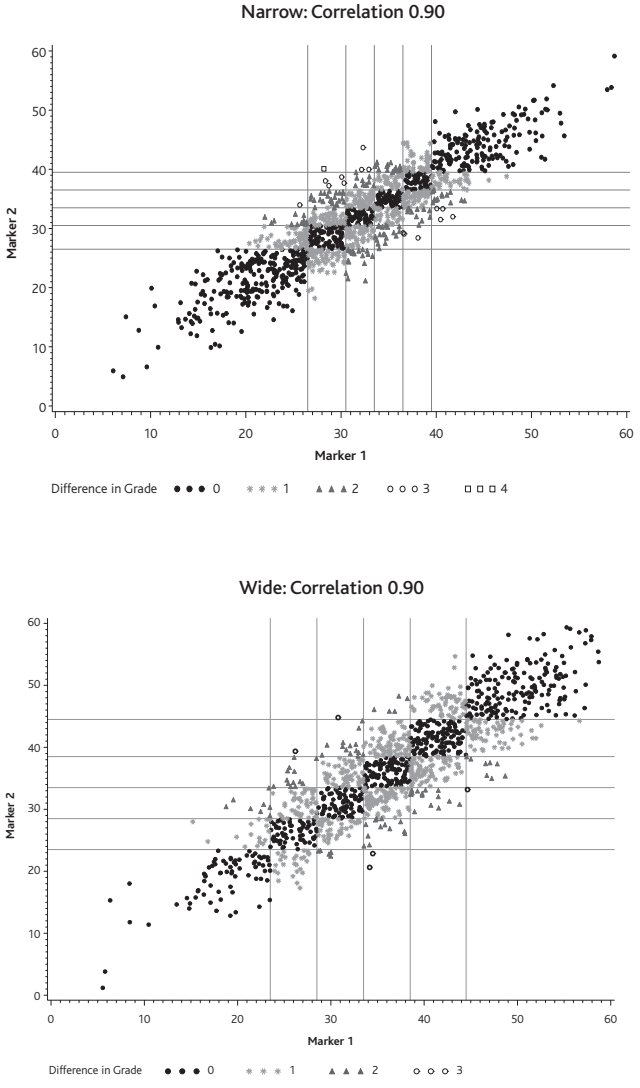


Figure 2 : Scatter plot of marks from two hypothetical markers (inter-marker correlation = 0.90)

It is clear that the impact on the proportion consistently classified of changes in the correlation coefficient between the two simulated markers was substantial. As expected, this fell with the level of correlation. For the narrow unit the percentage consistently classified fell from 86.5% at a correlation of 0.995 to 37.7% at a correlation of 0.7. For the wide unit the fall was slightly larger, from 88.1% to 35.6% consistently classified.

To demonstrate the levels of consistent classification visually, Figure 2 plots the marks from the two markers for both units, with a correlation of 0.90. Note that on the graphs the lines representing the boundaries have been set 0.5 marks below the actual boundaries, to show more clearly which mark points are in a particular grade and which are out.

We also looked at the classification consistency conditional on the *mark* given by one of the markers. This is the proportion of candidates on each mark (from marker x) given the same *grade* by marker y. This is best represented graphically, as shown in Figure 3.

These graphs demonstrate that for both units the levels of consistent classification fell considerably with marks on and around the grade boundaries (the vertical lines represent the boundaries). The peaks in the graphs are at marks in the middle of the boundaries. This is what we would expect, since for a mark on the grade boundary a difference of just one mark between the two markers (in one direction) is enough for inconsistent classification, whereas in the middle of the boundary a difference of two or three marks is necessary. It is worth noting that the differences between the peaks and troughs were much lower for low levels of correlation.

Severity of inconsistent classification

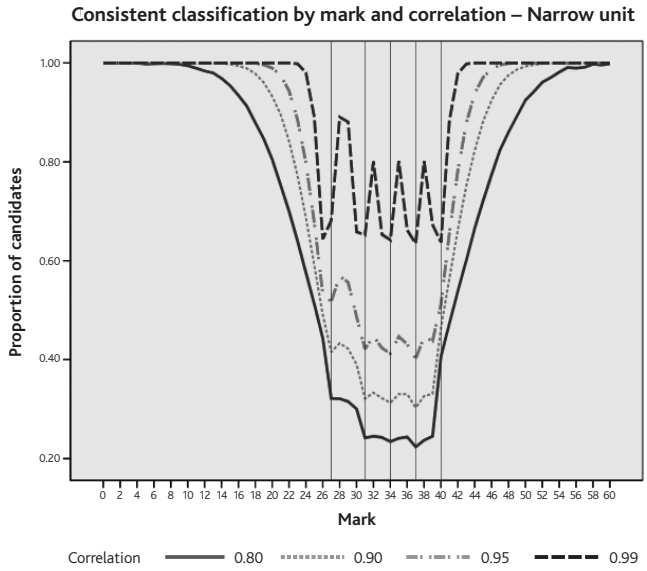
What the above indices do not take account of is the severity of the inconsistent classification – the proportions that were inconsistently classified by one grade, by two grades and so on. This is shown in Table 5 below:

Table 5 : Severity of inconsistent classification

	Correlation	*Proportion inconsistently classified by*					
		0 grades	*1 grade*	*2 grades*	*3 grades*	*4 grades*	*5 grades*
Narrow	0.995	0.865	0.135	<0.001	0	0	0
	0.99	0.809	0.191	<0.001	0	0	0
	0.95	0.617	0.341	0.042	0.002	<0.001	0
	0.9	0.523	0.363	0.099	0.014	<0.001	<0.001
	0.8	0.428	0.353	0.157	0.051	0.010	<0.001
	0.7	0.372	0.336	0.182	0.081	0.026	0.004
Wide	0.995	0.881	0.119	0	0	0	0
	0.99	0.832	0.168	<0.001	0	0	0
	0.95	0.637	0.349	0.014	<0.001	<0.001	0
	0.9	0.528	0.412	0.058	0.003	<0.001	<0.001
	0.8	0.418	0.427	0.132	0.022	0.002	<0.001
	0.7	0.356	0.414	0.175	0.047	0.007	<0.001

At correlations of 0.995 and 0.99 very nearly all of the candidates were classified within one grade for both units. At a correlation of 0.95 this was still the case, but the percentage inconsistently classified by one grade increased to over 30%. At a correlation of 0.90, around 11% of the candidates on the narrow unit and 6% of the candidates on the wide unit were inconsistently classified by two grades or more.

As with the proportion consistently classified we also produced graphs for the severity of inconsistent classification (by at least one, two or



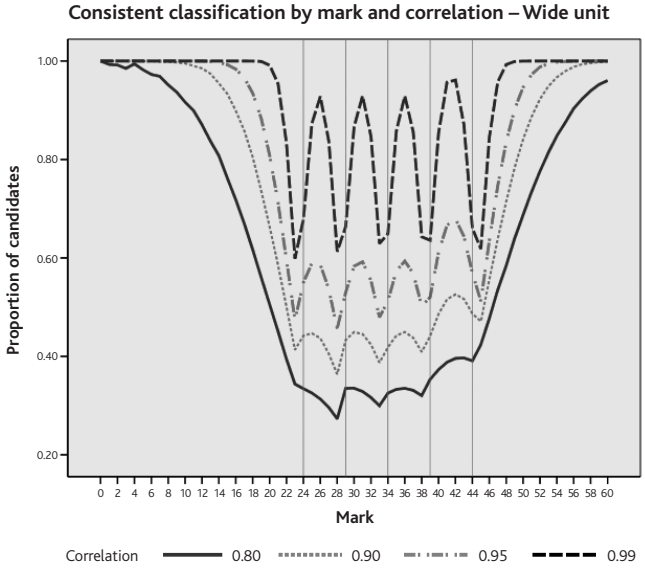
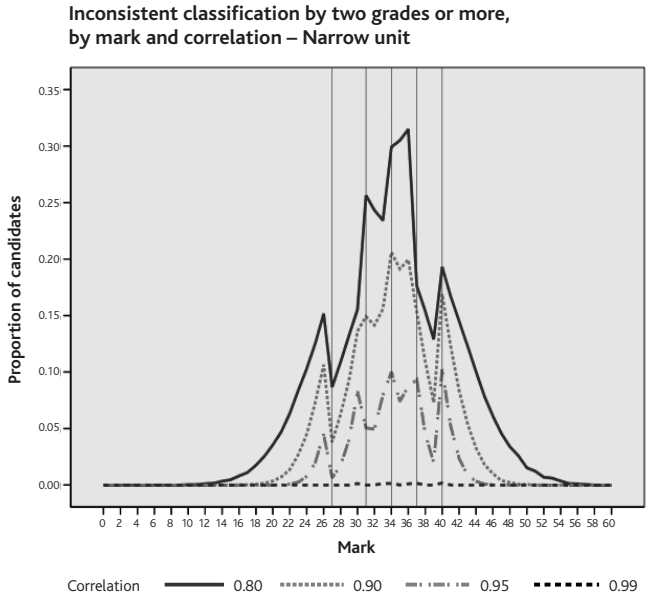


Figure 3 : Consistent classification by mark and correlation – Narrow Unit, Wide Unit[3]



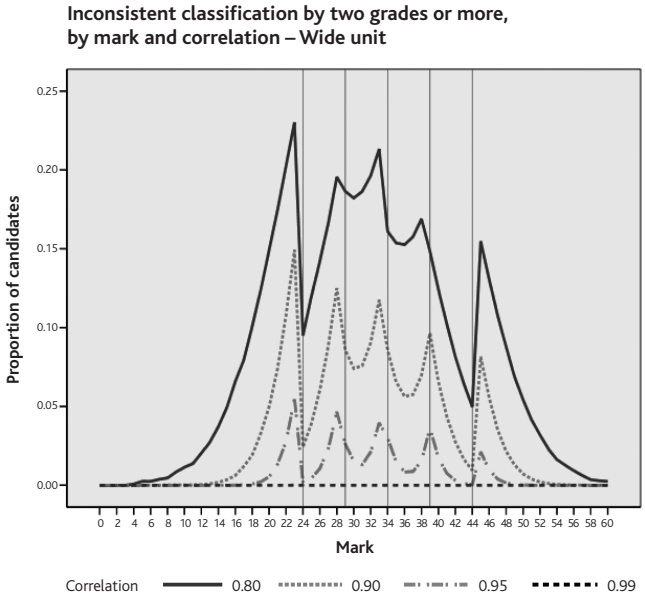


Figure 4: Inconsistent classification by two grades or more, by mark and correlation – Narrow Unit, Wide Unit

three grades), conditional on mark. Figure 4 above shows the proportion of candidates inconsistently classified by at least two grades, for both units.

As expected the graphs are generally the reverse of Figure 3, with the peaks on or around the boundaries; inconsistent classification is more likely on mark points close to the boundaries.

Differences between the units

The effect of altering the correlation between the two markers has been shown to be significant. A reduction in the correlation substantially reduced the proportion of candidates consistently classified and increased the severity of the inconsistent classification. We now consider the differences between the narrow and wide units.

Figure 5 shows the proportion consistently classified, and the proportion classified within one grade and within two grades for the narrow and wide units:

There was virtually no difference in terms of the proportion consistently classified, with the indices for the wide unit very slightly higher at high levels of correlation and the indices for the narrow unit very slightly higher at lower correlations. This was not what might have been anticipated since the wide unit had grade boundaries that were more spaced apart than the narrow unit and thus we expected less inconsistent classification. The reason for the similarity is the difference in the relative mark distributions of the units (see Table 3). The proportions in each grade were different and the standard deviation of the wider unit was larger (9.60 compared to 8.01) and so the distribution was also more spread out.

Where differences did occur between the subjects these were in the severity of the inconsistent classification. Figure 5 shows that the proportion of candidates classified within one grade and within two

3 We have only included four of the levels of correlation in these graphs so that they remain legible.

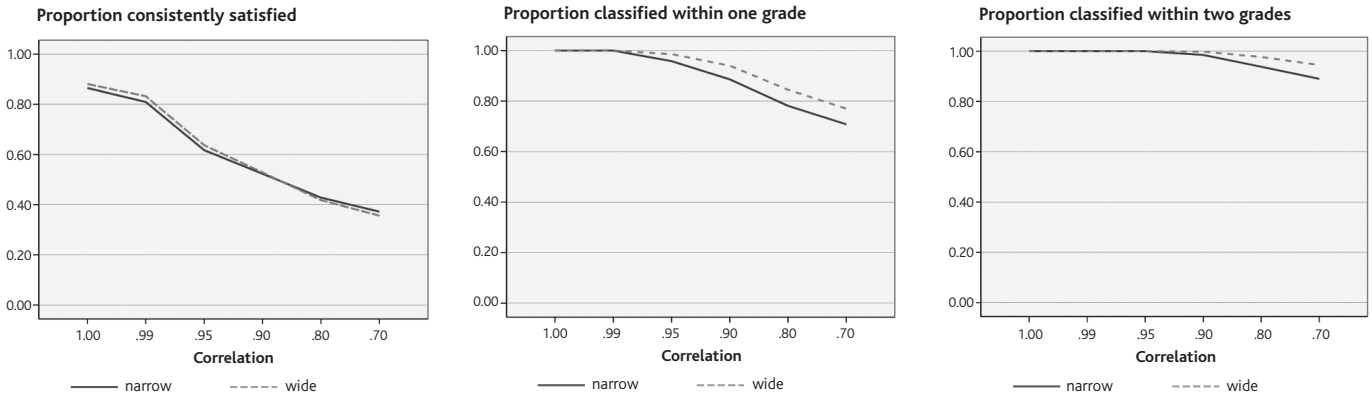


Figure 5: Comparison of levels of consistent classification between units

grades were both lower for the narrow unit. Hence the inconsistent classification in the narrow unit tended to be more severe. For example, in Table 5 we note that at a correlation of 0.90, 11.3% of candidates on the narrow unit were inconsistently classified by more than one grade compared with 6.1% of candidates on the wide unit. At a correlation of 0.80 these values were 21.8% and 15.6% respectively, at 0.7 they rose to 29.3% and 22.9%. Thus, in this example the overall effect of having more widely spaced grade boundaries was to reduce the severity, if not the amount of inconsistent classification.

Aggregation

The above analysis was at unit level. The candidate's overall grade at AS-level is of course based on the sum of the marks from each of three units. Thus the impact of inconsistent classification in any one unit is diluted by performance in the other two. However, inconsistent classification in the other units will also impact on the overall grade, and this could be compensatory to a candidate or it could make things worse. We used simulated data to investigate the impact of inconsistent classification on overall AS grade[4].

The 'wide' unit used above is an AS unit, and so we combined this with two other AS units in the same subject. There was some choice of units, but we chose the most popular units taken by those who took the original unit. We began by generating normally distributed data from the two markers on the first unit as above. We then used the real level of correlation in marks between each pair of pairs (units 1–2, 1–3 and 2–3) to simulate data for these other units, which were also normally distributed. Un-standardising each of these distributions (using the real means and standard deviations) gave a potential mark for each candidate on each unit. For the purposes of aggregation we then converted these to UMS[5]. Thus we had a mark for unit 1 by marker x (M_{1x}), a mark for unit one by marker y (M_{1y}), a mark for unit two (M_2) and a mark for unit three (M_3). For simplicity we started by assuming that there was only one marker on units two and three, so there was only one potential mark on each. The possible overall marks were thus:

$T1 = M_{1x} + M_2 + M3$

$T2 = M_{1y} + M_2 + M3$

From this the relative grades awarded under marker x and marker y, and thus the level of inconsistent classification, were estimated at each level of correlation.

We extended this analysis further by introducing inconsistent classification in unit two as well as unit one. So the totals we were interested in were:

$T_1 = M_{1x} + M_{2x} + M_3$

$T_3 = M_{1y} + M_{2y} + M_3$

T_1 is the total if units 1 and 2 were both marked by marker x and T_3 is the total if units 1 and 2 were both marked by marker y. We could then look at the proportion of candidates who would be consistently classified if not just one, but two of their units were marked by different examiners.

We used the same method as above, but just added another set of marks for the second unit and with a certain level of correlation in marks between marker x and marker y.

Finally, we introduced a second marker in the third unit, giving:

$T_1 = M_{1x} + M_{2x} + M_{3x}$

$T_4 = M_{1y} + M_{2y} + M_{3y}$

This time we were interested in the differences between T_1 and T_4 and the question became: what proportion of candidates would be consistently classified if all three of their units were marked by different examiners?

The results of the simulations are shown in Table 6 with the proportion consistently classified in terms of aggregated grade, compared with the consistent classification at unit level. The pairs of marks in each unit have the same correlation across all units.

Table 6 : Consistent classification in aggregated grade

Correlation	*P_0 (unit)*	*P_0 (aggregated, different markers unit 1)*	*P_0 (aggregated, different markers units 1 & 2)*	*P_0 (aggregated, different markers units 1, 2, 3)*
0.995	0.881	0.944	0.922	0.906
0.99	0.832	0.925	0.896	0.875
0.95	0.637	0.838	0.769	0.738
0.9	0.528	0.770	0.700	0.647
0.8	0.418	0.685	0.606	0.544
0.7	0.356	0.624	0.529	0.482

It is clear that the impact at aggregate level was much less than at unit level. As we suggested above, inconsistent classification in one unit is diluted when aggregated over the three units. In our simulation there was

4 This also applies to overall A-level grade, but it was simpler to use an AS level as an example, as this consists of only three units.

5 UMS=Uniform Mark Scale. See http://www.ocr.org.uk/learners/ums_results.html for a brief explanation. Note that in our example, the first unit had a maximum UMS of 120, whilst units 2 and 3 had a maximum of 90. Thus, the effect of misclassification of the first unit on aggregated grade is slightly greater than if all the units had equal weighting in terms of UMS.

also some 'averaging out' over the three units so that the potential levels of inconsistent classification at aggregate level were less than at unit level even if all three units were marked by different examiners. Thus at a correlation of 0.95 the potential inconsistent classification on one unit was 36.3%, compared to 26.2% at an aggregated level.

We have seen the effect of changes in the level of correlation between markers and the spread of the grade boundaries on the level of inconsistent classification, and also investigated the inconsistent classification at aggregate level. But what might this mean in reality for the number of pupils who would receive a different grade dependent on their marker? We estimated this using the levels of correlation from previous research.

There has been relatively little published research into marking reliability in UK exams. Murphy (1978, 1982) reported correlation coefficients of between 0.73 and 0.99 in 20 O-level and A-level subjects. As expected, subjects with more essay type questions such as English, History and Sociology tended to have lower levels of correlation than Maths and Physics papers, which are generally all short answer questions. Where more than one paper in each subject was investigated the aggregated marks generally correlated better than the marks on the individual papers. The correlations for the short answer questions varied from 0.98 to 1.00, whilst for the longer answer and essay type questions they varied between 0.73 and 0.98 with a mean correlation of 0.86.

More recently, Newton (1996) looked at correlations for Maths and English GCSEs. He reported correlations of above 0.99 for Maths and between 0.85 and 0.90 for English.

The two units in this research were quite different in that the paper for the narrow unit consisted of short answer questions and the paper for the wide unit was essay questions only. Thus if we arbitrarily allocate a correlation of 0.99 to the narrow unit and a correlation of 0.90 to the wide unit, we can estimate the potential levels of inconsistent classification. We should point out that this is not to suggest that these are the true levels of inconsistent classification, which cannot be known without blind double-marking, they are merely the levels that *might* exist, *if* the correlations were as stated. From Table 4, the percentage potentially inconsistently classified on the narrow unit was 19.1%, and the percentage for the wide unit was 47.2%. In other words, almost half of the students on the wide unit could potentially get a different grade dependent on the marker. Even on the narrow unit, where the level of inter-marker correlation is expected to be very high, up to one fifth of the candidates may be inconsistently classified.

The effect of aggregation would be to dilute the potential inconsistent classification. At the same level of correlation in the wide unit (0.90) 23% would be potentially inconsistently classified at aggregate (AS) level if one unit was marked by a different marker. This would increase to 35.3% if all three units were marked by different markers.

Conclusion

Since there is no such thing as a perfectly reliable test, there will always be a certain level of misclassification and/or inconsistent classification in tests and examinations. Exam boards go to great lengths to ensure that their procedures for marker monitoring, result checking and appeals allow all candidates the best chance of receiving the result that they deserve. However, the levels of misclassification/inconsistent classification are not well researched in relation to GCSEs and A-levels. Furthermore, it seems likely that the public underestimate the amount of measurement error that exists in these exams. If they were made aware of the true amount of error the level of trust in exam boards might be affected. Newton (2005) argues that while the level of trust may fall in the short term, there are many reasons why increased transparency about the extent of measurement error is desirable for students, policy makers, the public, and exam boards. His reasoning for this is 'it is crucial for those who use test and examination results to understand what kind of inferences can legitimately be drawn from them and what kind of inferences cannot' (Newton, 2005, p. 431). Because of the lack of understanding of measurement error, inferences might be drawn that cannot be justified. Whether or not this is the case, and whether it is likely that there will be more transparency in the future, we suggest that exam boards should be in a position to report an estimate of the amount of measurement error that exists in the qualifications they produce.

This article has presented the levels of inconsistent classification that *might* exist dependent on the marker used, based on simulating data in two A-level units, one with a particularly wide grade bandwidth and one with a narrow width. This should not be taken as evidence of the true levels of inconsistent classification in all A-level units, since each unit will have a different distribution of marks, a different grade bandwidth, and a different level of inter-marker correlation. However, this research does give an idea of the magnitude of the potential inconsistent classification, something that might come as a surprise to the general public.

Of course, there will always be a certain level of inconsistent classification since only completely objective tests will ever be free from measurement error attributable to markers. Further debate and investigation is needed into whether awarding bodies should routinely report estimates of these levels to the public. One approach would be to determine an acceptable level, and attempt to develop tests and train examiners so that this level can be attained. However, Newton (2005) argues that to define acceptable levels of accuracy is probably not realistic given the different natures of exams and the trade-offs between 'technical characteristics and pragmatic constraints'.

Alternatively, given that there will always be a level of inconsistent classification, more than one grade could be reported (Please, 1971) or confidence intervals could be reported on the mark given (Newton, 2003). Please suggested reporting grades in the following clusters; A/B, A/B/C, B/C/D, C/D/E, D/E/O, E/O/F and O/F. However, as he himself stated, this could lead to people treating A/B as the top grade, A/B/C as the next and so on, ignoring the implication that the candidate's true grade could be any of those in the group. The idea of confidence intervals is to report a range of marks within which we are almost certain the candidate's observed score will lie for a given true score. This method would give an idea of how much reliance we should put on exam results as an accurate summary of a candidate's skills in a particular area, and would therefore mean it is less likely that the results would be used to make unrealistic inferences.

Another idea would be to report for each grade an estimate of the proportion of candidates with that grade who might have received a higher or lower grade if another marker had marked the paper. As an example, Table 7 shows this for the narrow unit if the inter-marker correlation was 0.90.

Thus 27.2% of the grade B candidates might have got a higher grade from a different marker, and 40.9% might have got a lower grade. This

Table 7: Proportion of candidates getting a higher or lower grade if marked by a different marker

Observed grade	*Proportion Higher*	*Proportion Lower*
A	0.000	0.265
B	0.272	0.409
C	0.302	0.374
D	0.336	0.339
E	0.330	0.255
U	0.227	0.000

would be a relatively easy way of understanding how much reliance should be put on the results given. A table like Table 7 is a more informative version of a reliability coefficient. Like a reliability coefficient it is not a fixed property of the test, but depends on the distribution of scores, the grade bandwidth and (in this case) the inter-marker correlation. The proportions cannot be interpreted as probabilities for individual candidates, however, because this would depend on how close the individual was to the grade boundary. The proportions apply to the grade scale as a whole.

Finally, some limitations of this study should be mentioned. First, we mainly looked at levels of inconsistent classification in one unit only. In reality this may not be as important to candidates, as we have shown the effect is almost certain to be diluted when aggregating over the three units of AS. This would be even more the case when aggregating over six units of A-level. Arguably, it is at the aggregate level that any inconsistent classification is particularly serious: for example, when grades are used to create point scores for university selection. Secondly, it may be that using a normal distribution to simulate the data is not the ideal method. For instance, having to truncate the distribution at zero and the maximum mark meant losing some of the data, and may have slightly distorted the distribution. It may be that other distributions would better match the distribution of the data in reality, such as the beta binomial (see Livingston and Lewis, 1995; Lee *et al*., 2000). Finally, this research only considered inconsistent classification arising from differences in correlation between markers' scores, not differences between markers in severity or bias. Future research could address some of these issues, and widen the scope to other assessments, such as GCSEs or admissions tests.

References

Bramley, T. (2007). Quantifying marker agreement: terminology, statistics and issues. *Research Matters: A Cambridge Assessment Publication*, **4**, 22–28.

Cresswell, M. (1986). Examination grades: How many should there be? *British Educational Research Journal*, **12**, 1, 37–54.

Lee, W-C., Hanson, B.A. & Brennan, R.L. (2000). Procedures for computing classification consistency and accuracy indices with multiple categories. *ACT Research Report Series*. Available online at http://www.act.org/research/reports/pdf/ACT_RR2000-10.pdf (accessed 23 October 2006)

Livingston, S.A. & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement*, **32**, 2, 179–198.

Lord, F.M. & Novick, M.R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Murphy, R.J.L. (1978). Reliability of marking in eight GCE examinations. *British Journal of Educational Psychology*, **48**, 196–200.

Murphy, R.J.L. (1982). A further report of investigations into the reliability of marking of GCE examinations. *British Journal of Educational Psychology*, **52**, 58–63.

Newton, P.E. (1996). The reliability of marking of General Certificate of Secondary Education scripts: Mathematics and English. *British Educational Research Journal*, **22**, 4, 405–420.

Newton, P.E. (2003). The defensibility of national curriculum assessment in England. *Research Papers in Education*, **18**, 2, 101–127.

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

Please, N.W. (1971). Estimation of the proportion of examination candidates who are wrongly graded. *British Journal of Mathematical and Statistical Psychology*, **24**, 230–238.

Vidal Rodeiro, C.L. (2007). Agreement between outcomes from different double marking models. *Research Matters: A Cambridge Assessment Publication*, **4**, 28–34.

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

EXAMINATIONS RESEARCH

Statistical Reports: Patterns of GCSE and A-level uptake

Joanne Emery and Carmen L. Vidal Rodeiro Research Division

Two new statistical reports have been added to the 'Statistics Reports' series on the Cambridge Assessment website (http://www.cambridgeassessment.org.uk/ca/Our_Services/Research/Statistical_Reports):

Statistics Report Series No. 4: *Uptake of GCSE subjects 2000–2006*

Statistics Report Series No. 5: *Uptake of GCE A-Level subjects in England 2006*

Data for these reports were extracted from the 16+/18+ databases. These databases are compiled for the Department for Children, Schools and Families (DCSF) from data supplied by all the awarding bodies in England. They contain background details and national examination data for all candidates who have their 16th, 17th and 18th birthdays in a particular school year. Candidates are allocated a unique number that remains the same throughout their Key Stage tests, allowing matching of examination data for longitudinal investigations. Records are present only if the candidate has sat an examination in a particular subject, not just attended classes.

This brief article outlines some of the results from both reports.

Uptake of GCSE subjects 2000–2006

There were a total of 561,407 students that attempted at least one GCSE examination in 2000. This number increased 12% to reach 629,523 students in 2006. The average number of GCSEs taken by candidates in

the database was 8.36 in 2000 and 7.95 in 2006. This slight decline might be due to the increase in flexibility in GCSE studies, with new applied options for traditionally academic subjects (recorded as different qualifications), changes in the National Curriculum requirements, increased use of entry level qualifications or new 'hybrid' GCSEs that allow students to study on either academic or applied tracks.

An example of the results in the report is described here: the uptake of GCSE science subjects. In Statistics Report Series No. 4, similar analyses for almost all GCSE subjects are available.

The uptake of the separate sciences (biology, chemistry and physics) increased slightly from 2000 to 2006 but, on the other hand, the uptake of the double award in science fell almost 8 percentage points from 2000 to 2006. Girls were less likely to take the separate sciences at GCSE (which will limit their opportunities to progress onto science-based advanced level study). The uptake of biology, chemistry and physics was higher for the higher attaining students. This may explain why the percentages of students entered for the single and double award science courses was lower for the high attaining group compared to the medium and low attaining groups. The uptake of the separate sciences was much higher in independent and grammar schools than in comprehensive and secondary modern schools. With regard to the science double award, the uptake increased in independent schools (around 11 percentage points) but decreased in other types of schools. The uptake of the science single award increased only in comprehensive schools.

Other variables, such as the school gender, school boarding status or the characteristics of the neighbourhood in which the school is situated, were considered in this report and the uptake of the science subjects in 2006, according to candidates' school gender and various of the school neighbourhood variables, is presented in Table 1. Neighbourhood variables were downloaded from the Office of National Statistics Census 2004 data and were matched to the examination data according to the postcode of the school.

Table 1 : Uptake of GCSE science subjects in 2006 (percentages of students taking GCSEs)

		Biology	*Physics*	*Chemistry*	*Science: double award*	*Science: single award*
School gender	Boys	25.2	24.9	24.8	58.8	8.1
	Girls	13.2	12.4	12.7	70.7	9.3
	Mixed	6.5	6.2	6.3	71.0	11.6
Urban/rural indicator	Urban	8.0	7.7	7.8	69.1	11.6
	Town	6.9	6.7	6.7	78.3	8.9
	Village	9.5	8.8	9.0	74.6	9.6
Income deprivation affecting children	Bottom (lowest deprivation)	9.7	9.4	9.5	72.6	9.4
	Middle	7.1	6.8	6.8	70.1	12.2
	Top	5.2	4.8	5.0	64.5	14.1
% working-aged people with no qualifications	Bottom (lowest deprivation)	12.5	12.1	12.2	71.5	8.3
	Middle	7.1	6.8	6.8	72.5	10.8
	Top	4.4	4.1	4.2	66.2	14.7

Results for classifications based, for example, on school boarding status, multiple deprivation, employment rate and the percentage of people with Level 4/5 qualifications, are available in Statistics Report Series No. 4.

Uptake of GCE A-level subjects in England 2006

A total of 223,710 students in England attempted at least one A-level examination in 2006 (an increase of 7,897 students, or 3.7%, from the previous year). This figure equals less than a third of the number taking GCSE examinations in 2006. The modal number of A-level examinations taken was 3 (representing 49% of all candidates), followed by 4 (24% of candidates). If General Studies is excluded then 63% of all candidates attempted only 3 A-level examinations. These figures are similar to those reported previously for 2002 to 2005 in Statistics Report Series No. 3.

Statistics Report Series No. 5 lists the 30 most popular A-level examinations taken in 2006 and tabulates the percentages of candidates taking each of these subjects according to their school type, school gender and various school neighbourhood factors (mostly indicators of deprivation). The number of subjects and 'LEP' subjects taken (subjects listed by the University of Cambridge as providing 'less effective preparation' for their undergraduate courses) are also tabulated by these factors. The top 30 combinations of 3 or more A-level subjects is also presented.

The uptake of A-level science subjects and maths is presented in Table 2, categorised by candidates' school gender and a selection of school neighbourhood variables. Continuous variables (such as the percentage of working-aged people with no qualifications) were divided into three equal-sized groups using percentile values. The groupings here do not represent England as a whole because those from disadvantaged backgrounds are less likely to take A-levels. The full report additionally contains classifications based on school type and boarding status, estimates of neighbourhood income and the percentage of people with Level 4/5 qualifications.

Table 2 : Uptake of A-level science subjects and maths in 2006 (percentages of students taking A-levels)

		Biology	*Chemistry*	*Physics*	*Maths*
School gender	Girls' Schools	26.1	20.4	7.2	22.1
	Girls in Mixed Schools	17.7	10.5	2.8	11.9
	Boys' Schools	22.4	22.5	20.0	36.2
	Boys in Mixed Schools	15.4	14.4	16.0	24.8
Urban/rural indicator	Urban	17.8	13.8	9.3	19.6
	Town	18.8	13.0	10.8	18.2
	Village	19.1	15.4	11.8	22.6
Income deprivation affecting children	Bottom (lowest deprivation)	18.5	14.4	10.4	20.7
	Middle	17.5	13.2	9.6	19.6
	Top	17.8	13.7	8.6	18.5
% working-aged people with no qualifications	Bottom (lowest deprivation)	19.3	15.8	10.5	22.4
	Middle	17.5	13.3	10.0	19.3
	Top	17.0	12.3	8.1	17.0

The uptake of A-level science subjects and maths was higher in girls' schools than for girls in mixed schools. The uptake of English Literature and foreign languages was higher in boys' schools than for boys in mixed schools. However, single-sex schools are much more likely to be independent or grammar schools and these factors themselves were associated with higher uptakes of these subjects (some of the complexities of interpreting the examination results for single sex schools

are discussed in a recent paper by a former Cambridge Assessment research officer (Malacova, 2007). Students attending schools and colleges in areas of higher deprivation were more likely to take fewer A-levels and more likely to take a higher number of LEP subjects. This will limit their opportunities to apply to courses at the University of Cambridge (a student will normally need to offer at least two non-LEP subjects). However, the differences are relatively small and did not take into account their previous attainment at GCSE.

Reference

Malacova E. (2007). Effects of single-sex education on progress in GCSE. *Oxford Review of Education*, **32**, 2, 223–259.

QUALITY AND STANDARDS

The OCR Operational Research Team

Elizabeth Gray OCR

To those within OCR (Oxford, Cambridge and RSA Examinations) the Operational Research Team (ORT) provides a constant source of advice, data and statistical support on all technical matters; to those outside OCR their work is largely unknown. This short sketch is an introduction to the main areas of interest of the team and its involvement in the life of OCR.

The outline will start, since at the time of writing the summer awarding series for general qualifications has just been completed, at the end, with the support provided to Awarding Committees and, crucially, Subject Officers and Chairs of Examiners. General assessments are becoming increasingly technical and the use of prior attainment measures to predict outcomes for both GCE and GCSE examinations requires technical manipulation of the highest order. Modelling aggregation (subject level) outcomes in unit based assessments is an essential part of awarding preparation and one which would cause problems were EPS (Examinations Processing System) to be used. In addition, where new subjects are awarded, additional data are provided to help with decision making. The awarding session also brings with it malpractice cases and the ORT supports the malpractice process and helps with the running of malpractice and appeals committees.

This work, though very intense, actually only represents a relatively small part of the ORT's programme. Vocational qualifications are awarded on a more continuous basis than general qualifications and again the ORT provides support for that process. This may, for some assessments, include producing question papers from a library of questions using complex statistical techniques to ensure standards are maintained.

New qualifications provide a source for much of the ORT's work and technical advice is sought regarding the assessment structures and marking regimes. When new specifications are proposed, for example the four unit A-levels, preparatory work is done to gauge the effect of the new assessment structure – in this example the effect of the decrease in the number of units on specification grade distributions. The outcomes from the work will again feed into awarding committees, and new developments, to aid the decision making process. When the issue is likely to affect all awarding bodies, for example the A* at GCE, then the research will be in collaboration with the Joint Council for Qualifications (JCQ). Indeed, many of the investigations undertaken by the ORT are at the behest of the Qualifications and Curriculum Authority (QCA) or the JCQ and contribute to a pool of knowledge shared by all awarding bodies.

QCA often want new qualifications to be trialled or piloted, as is the case for functional skills, and these trials/pilots have to be evaluated both for our own requirement and also for QCA as part of the pilot contract. The ORT has a standing programme of such evaluations which focuses mainly on the innovatory aspects of the trial or pilot and equivalence with existing qualifications. It was on QCA's behalf that a 'Stretch and Challenge' trial was conducted recently on new A-level assessments. This initiative was led by the ORT who will also be analysing the data once the scripts have been marked. The results of the analysis of this trial will be shared with all awarding bodies and QCA at a seminar in November 2007.

National Curriculum testing is now declining, but OCR took over that responsibility from the Assessment Research and Development division (ARD) of Cambridge Assessment in September 2005. This has led to a build-up of expertise in item level analysis which will stand OCR in good stead in the new e-environment. Collaboration across business streams on electronic script management (ESM) research has also enhanced knowledge in that area which can now be put to practical use.

A new member of the team, recruited in March 2006, has allowed more investigation into Malpractice, Appeals and Result Enquiries to take place. By identifying those subjects which attract the greatest number of events and changes arising from those events, research into underlying root causes can feed into specification development and strategies for improving marking reliability.

The quality of marking is always of concern, so much so that an internal OCR committee has been set up to consider the issues and identify investigations to be carried out by the ORT. Led by an ORT member, this committee also has presentations given by ARD members when their research relates to marking issues when the practical application of the research findings is considered.

When time permits, some of the issues raised by straightforward technical investigations lead to more detailed research. For example, as part of the continuous statistical monitoring of awarding decisions, research into awarding judgements showed that awarders cannot easily differentiate scripts which are only 2 or 3 marks apart. This finding lends support to the current awarding process where a zone of marks is defined by judgement of scripts and statistical considerations help to identify the final boundary mark within that zone.

The more OCR knows and understands about its processes the fewer errors are likely to be made and although it is the ORT's role to anticipate assessment issues and provide information to mitigate them, there is no doubt that trouble shooting is also required. In order to reduce this, the ORT is heavily involved in training Subject Officers and Chairs in all

technical aspects of the assessments which OCR offers and in the understanding of the statistical data which are and can be provided for all stages of the assessment process.

The ORT in its current form has been in existence since 2004 (there has been a small in-house research facility since the creation of OCR). The team is seven strong, six of whom are based in Cambridge, and is headed by an Assistant Director. Four of the team served their apprenticeship in the Research and Evaluation division (the fore-runner of the Research Division) and can be fairly described as having over a quarter of a century of research experience between them. The somewhat narrower focus of their work is essential given the immediacy of the applicability of any results whether in the development, question setting and marking or awarding and post-awarding stages of the assessment process.

Research News

Conferences and seminars

International Association for Educational Assessment

The 33rd annual conference of the International Association for Educational Assessment (IAEA) took place in Baku, Azerbaijan in September. The main purpose of the IAEA is to assist educational agencies in the development and appropriate application of educational assessment techniques to improve the quality of education. This year's conference was hosted by The State Students Admission Commission of the Republic of Azerbaijan and was attended by 343 delegates from 43 countries. The conference theme was: *The interpendence of national assessment systems and education standards*. The Assessment Research and Development (ARD) Division presented six papers – including a paper on *Promoting educational quality through national assessment systems* by ARD Group Director, Tim Oates and Research Division Director, Sylvia Green.

British Educational Research Association

In September, 10 colleagues from the Research Division attended the British Educational Research Association (BERA) annual conference at the Institute of Education, University of London. A total of 11 papers were presented, reporting on a wide range of issues from vocational grading to how question features relate to marking accuracy.

Royal Statistical Society

Carmen Vidal Rodeiro attended the Royal Statistical Society conference in York in July. The theme for 2007 was *Statistics and public policy making*. Around 200 presentations were given and specific themes included crime, education, trust in statistics, statistical legislation, and the way statistics contribute to policy.

European Conference on Educational Research

Jackie Greatorex presented a paper – *Exploring how the cognitive strategies used to mark examination questions relate to the efficacy of examiner training* – at the European Conference on Educational Research (ECER) in September at the University of Ghent. ECER is the annual meeting of the European Educational Research Association (EERA) which constitutes the British Educational Research Association (BERA) and similar organisations of the other European nations. Over 1000 delegates attended, mainly from Europe but some from farther afield e.g. University of Zimbabwe, University of Japan and University of New York. About 880 papers were presented, and 30 symposia and 10 workshops took place.

Journal of Vocational Education and Training conference

Martin Johnson attended the Journal of Vocational Education and Training 7th International Conference, University of Oxford, in July and presented a paper entitled: *Grading, motivation and vocational assessment*. He also presented a paper by Nadežda Novaković on *The influence of statistical data on panellists' decisions at Angoff awarding meetings*.

Association for Educational Assessment – Europe

In November three colleagues from the Assessment Research and Development Division attended the annual conference of AEA-Europe in Stockholm and presented papers. The theme of the conference was *Assessment for educational quality*.

Cambridge Assessment Conference

The 3rd Cambridge Assessment Conference took place at Robinson College, University of Cambridge, in October. The conference theme was *e-Assessment and its impact on education* and took a broad view of the potential of new technologies to improve assessment, with the purpose of identifying and promoting those innovations that will create valid assessments as well as educational benefits. The main speakers were Professor Andrew Pollard from the ESRC Teaching and Learning Research Programme, Institute of Education, University of London and Professor Richard Kimbell, Goldsmiths, University of London. Twelve discussion seminars enabled delegates to debate issues on a range of subjects within the main conference theme.

Research Division seminar

In October a research seminar was held at Cambridge Assessment entitled: *How can qualitative research methods inform our view of assessment?* Professor Harry Torrance and Dr Helen Colley of the Education and Social Research Institute, Manchester Metropolitan University, and Martin Johnson of the Research Division gave presentations exploring the potential of qualitative research methods for understanding aspects of assessment that are difficult to capture through quantitative surveys and measurement. The presentations drew on projects funded by the ESRC and the LSDA to illustrate how a qualitative approach can inform our view of assessment.

A date for your diary

As part of our 150th anniversary celebrations in 2008, Cambridge Assessment will host the 34th annual conference of the International Association for Educational Assessment (IAEA). The annual IAEA conference is recognised as a major event in assessment, bringing together leading assessment and education experts from across the world.

Date – Sunday 7 to Friday 12 September 2008.

Venue – the conference will take place in Cambridge, UK. The main conference sessions will be held at Robinson College, Cambridge University's newest college.

Theme – *Re-interpreting assessment: society, measurement and meaning*. Sub-themes will range from *Emerging trends and perspectives in assessment* to *Equality issues in assessment*.

Keynote speakers – Professor Robert J. Mislevy, University of Maryland, and Professor Dylan Wiliam, Institute of Education, University of London.

Registration – registration and 'call for papers' will open on 14 January 2008.

Further information can be found at:
www.iaea2008.cambridgeassessment.org.uk/.

Cambridge Assessment
1 Hills Road
Cambridge
CB1 2EU

Tel: 01223 553854
Fax: 01223 552700
Email: ResearchProgrammes@cambridgeassessment.org.uk
http://www.cambridgeassessment.org.uk

ISSN: 1755–6031