to allow for differences in demand, the point about the UMS conversion is that this differential has been allowed for. Looked at from a raw mark perspective, if specification uniform mark grade boundaries are allowed to fluctuate (but not unit conversions), then the relationship of raw unit boundaries to that final total will vary. Even calculating a regression allowance of UMS marks would lead to year-on-year anomalies because candidates on what were ostensibly equivalent marks could achieve different grades purely because of the company they keep even though much of their assessment might be common.

Conclusions

This article has attempted to explain the underlying rationale for the employment of uniform marks: their conception, their computations; and their effect on a range of aggregations. The principal motivation for using the uniform mark scale relates to the structure of regulation for GCE specifications and of choice for GCSE development unit based.

The relative strengths and shortcomings of using uniform marks for unitised schemes of assessment are both multiform and various. Unitised schemes are flexible, enhance overall performance (although some would say unfairly because of the provision for re-sits) and enable weaker candidates to show what they know, understand and can do because the learning approach is both incremental and developmental: learners have greater control regarding choice of assessment without undue reliance on terminal assessment. Unitised assessments are manageable, formative and can be delivered at the point of learning within the programme of study. Additionally, GCE and GCSE are similar in basic structure with units employing credit ratings which have the potential to be used in a National Qualifications Framework and as part of the Additional and Specialised Learning in the Diploma.

Conversely, there is a prevailing belief that unitisation can lead to increased testing and, therefore, to a concomitant increase in the burden of assessment. More disturbingly, there exists a public perception that unitised schemes are easier, largely due to the re-sit policy. From a cognitive maturation perspective, it is also held that some candidates who take unitised assessments may forget that part of the curriculum very readily. This has led to synoptic assessment in GCE specifications and terminal rules for the new GCSE developments.[4] In terms of their interpretation, evidence would suggest that centres find it difficult to read and comprehend UMS data. We have seen that there are problems when there are discontinuities in the conversion rates which have led to the generation of some additional rules to maintain conversion parity.

Whatever the arguments, the UMS system has stood the test of time (it was first introduced as a mechanism for aggregating GCE specifications in the late 1980s) and, with the modifications described, seems to work well. There are concerns that with the new A levels and the introduction of 'stretch and challenge' questions it will be difficult to target grades as precisely as is achieved with the current GCEs with the inevitable consequences of low grade A and, possibly, E boundaries. GCE A* is another complication because its achievement is crucially dependent on the amount of capping there is in the specification. But until another, more effective, system is devised for aggregation, uniform marks are likely to remain.

References

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T. and McNamara, T. (1999). *Dictionary of Language Testing, Studies in Language Testing 7*. Cambridge: UCLES and Cambridge University Press.

Greatorex, J. and Malacova, E. (2006). Can different teaching strategies or methods of preparing pupils lead to greater improvements from GCSE to A level performance? *Research Papers in Education*, **21**, 3, 255–294.

Qualifications and Curriculum Authority (2000). *Arrangements for the statutory regulations of external qualifications in England, Wales and Northern Ireland*. London: QCA.

Qualifications and Curriculum Authority (2007). *GCSE, GCE, GNVQ and AEA Code of Practice*. London: QCA.

Patrick, H. (2003). Synoptic Assessment: A report for QCA. Available at http://www.ofqual.gov.uk/files/synoptic_assessment-_report_for_qca_pdf_05_1620.pdf

Pollitt, A., Ahmed, A., and Crisp, V. (2007). The demands of examination syllabuses and question papers. In: P. Newton, J-A Baird, H. Goldstein, H. Patrick, and P. Tymms (Eds.), *Techniques for Monitoring the Comparability of Examination Standards*. London: Qualifications and Curriculum Authority.

Stevens, J. (2002). The demands of synoptic assessment in the new English literature A level. *The Use of English*, **53**, 2, 97–107.

Thomson, D. G. (1992). *Grading Modular Curricula*. Cambridge: Midland Examining Group.

4 QCA has defined synoptic assessment as follows (QCA, 2000): A form of assessment which tests candidates' understanding of the connections between the different elements of a subject. See also Patrick, H (2003) and Greatorex and Malacova, (2006).

CAMBRIDGE INTERNATIONAL EXAMINATIONS

The CIE Research Agenda

Stuart Shaw CIE

Introduction

Cambridge Assessment has long devoted attention to assessment research. As part of its on-going commitment to examination quality, Cambridge International Examinations (CIE) has developed and established a unit dedicated to research. Although small, the team is responsible for a variety of research activities ranging from routine operational procedures in support of the quality of assessment processes to more full-scale experimental investigations whose purpose is to inform and improve on those operational procedures.

The research unit is responsible for three main areas of activity:

- **Routine operational analysis** concerning the management cycle of all CIE assessments, including the examination production, conduct, marking and awarding, and post-examination appraisal.

- **Instrumental research** concerning trials, projects and studies which are designed to inform the operational activities but which could not ordinarily be addressed as part of routine operational work. This might involve work related to the validation of existing or proposed syllabuses and the decision to revise certain features of an examination prior to its implementation in the live operational context; investigation of construct validity in selected syllabuses; comparability of standards across examination boards; or the impact on traditional assessment practice and marking reliability of translating from paper-based to screen marking.
- In addition to a planned research programme of activities it continues to be important that the research unit is also able to provide **reactive research** capacity when necessary. This is essential in an ever-changing and demanding operational environment.

In order to enhance the fairness of CIE examinations it is crucial that an agenda is created which establishes the necessary requirements for sound testing practice and which embodies the assessment arguments which underpin the examinations offered. Given the importance of meeting the need of high standards of quality and fairness, a range of key assessment considerations have been identified which currently underpin the research agenda. These are organised into six strands of activity: reliability and validity; comparability and standards; new technology; development of new CIE products and procedures; marking and awarding; and commissioned research. Each strand contributes to the achievement of maximum examination 'usefulness' in relation to intended contexts of use, that is, usefulness in fulfilling an intended purpose.

1. Reliability and validity

Traditionally, the quality of a test is assessed in relation to two key qualities: reliability and validity. The pursuit of high reliability is a continuing goal of CIE test construction. In the context of testing, reliability denotes dependability. In the sense that a test is deemed reliable, it can be depended on to produce very similar results in repeated uses. Thus reliability relates to replicability (stability and consistency), precision and overall test fairness. However, a test exhibiting high reliability may not necessarily measure the underlying skill of interest. This is where validity assumes importance – does the test measure what it is supposed to measure? If it does it is said to have validity. Cambridge Assessment treats validity (and validation processes) as a pervasive concern which permeates all of its work on the design and operation of assessment systems. Cambridge Assessment also acknowledges the inter-relationship between validity and reliability: where validity is poor, reliability in the assessment is of little value. If reliability is poor, that is, if the test results lack stability, then validity is compromised.

As a high stakes examination provider, CIE is committed to providing appropriate evidence for the validity and reliability of its range of assessments. Examples of research projects undertaken in this area include:

- **Validating revised and proposed new CIE syllabuses:** adoption of a new syllabus or the revision of an existing one requires that appropriate validation studies are conducted before the assessment can be implemented in a live context.
- **Predictive validity research into student performance in first year undergraduate studies:** since the main purpose of a test is to provide information about likely behaviour in the real world, prediction of criterion performance is basic to test validation and essential for the credibility of CIE assessments. Such studies are extremely useful when liaising with universities. For example, a claim made about CIE A Levels and the new Pre-U is that they provide an excellent preparation for university study. Predictive validity studies provide proof absolute of these claims. Such research is also helpful in resolving issues of equivalence with local qualifications in specific contexts.
- **Articulating construct(s) underpinning CIE assessments:** CIE are presently exploring more effective ways to demonstrate and share how it is attempting to meet the demands of construct validity in its range of assessments. This is achieved through a coherent programme for test development, validation and research to support claims about the validity of the interpretation of its qualifications results and uses, and by demonstrating evidence of the context, cognitive and scoring validity of the test tasks it provides.

2. Comparability and standards

Comparability, the application of the same standard across different examinations, remains a key concern in relation to the provision of large-scale educational assessments in England. CIE are committed to a rolling programme of work to ensure the equivalence of standards of similar qualifications across different awarding bodies, both national and international. Comparability studies embrace a range of CIE assessments including IGCSE, International A Level and O Level.

Traditionally, inter-examination board comparability studies have focused on the notion of 'score equivalences', that is, how the grades from each examination relate to one another, and the ways in which the examinations can be thought of as being 'comparable'. The research team has recently broadened the scope of what constitutes a comparability study by taking into account specific educational contexts.

By reviewing a range of comparability techniques, CIE are contributing to the development of a uniform approach to comparability studies and methodology across Cambridge Assessment, an approach which will constitute the basis of a future comparability programme.

3. New technology

With the development of new technologies, many tests that have previously been available and assessed on paper are being adapted and marked in more innovative ways. The ability for Cambridge Assessment examiners to mark a script from an on-screen image is provided via Scoris® software. Scoris® displays digital images of the scripts on-line through a web-based system and enables examiners' marks and notations to be recorded and the marks automatically returned to Cambridge Assessment. In transferring from one assessment medium to another, however, it is crucial to ascertain the extent to which the new medium may alter the nature of traditional assessment practice or affect marking reliability. As a result, CIE has expended considerable time, effort and resource in order to determine in exactly what circumstances on-screen marking is both valid and reliable.

The research team have been engaged in a series of on-screen essay marking trials (Checkpoint English and the General Paper), which have attempted to investigate marker reliability, construct validity and whether factors such as annotation and navigation differentially influence examiner performance across marking modes. The marking pilots had sought to ascertain whether examiners make qualitatively different assessments when marking the same piece of writing but

through a different medium. The trials have influenced the decision to move to online marking of Checkpoint English and have generated a number of recommendations for improving the current functionality of the software. These trials have highlighted the challenge of maximising ease of marking without compromising assessment validity.

Other areas of research include investigations into the feasibility of remotely standardising examiners in a Scoris® environment – the marking application provides the potential for an on-line mechanism for more effective virtual examiner co-ordination; and, exploring how the availability of item-level data generated by Scoris® – marked CIE assessments can be utilised to inform existing grading procedures.

4. Development of new CIE products and procedures

Considerable research attention is given to the introduction of new CIE products. One new assessment, the Cambridge Pre-U, is a post-16 qualification that prepares students with the skills and knowledge they need to make a success of their subsequent studies at university. It is part of Ofqual's remit (formerly, QCA's remit) to monitor all qualifications it accredits, including the Pre-U. CIE is keen to co-operate as fully as possible in this process. To do this, CIE are now starting to give consideration to any issues surrounding the monitoring procedures that may emerge. To this end, a working liaison is being developed between the respective research departments of Cambridge Assessment (CIE and ARD) and Ofqual. Being able to determine monitoring requirements throughout the monitoring process, and in advance of that process, will facilitate the passage of any pertinent documentation (such as academic papers and the findings from various research studies); and, help determine any appropriate trial methodologies necessary for satisfying the requirements of the monitoring process.

In addition to the monitoring and evaluation work, several UK universities have offered to assist CIE with research on the Pre-U. One suggestion is that the universities set the Cambridge Pre-U papers to their new intake in October in relevant subjects. If data from the students are collected on their A Level results, the papers and Pre-U results could comprise part of the Pre-U standards setting exercise. This exercise would take place in October 2008 and 2009.

Another area of interest relates to the *Content and Language Integrated Learning* (CLIL) project. CLIL is defined as an approach in which a foreign language is used as a tool in the learning of a non-language subject in which both language and the subject have a joint role. CLIL programmes currently operate in a range of different linguistic contexts and are, therefore, open to a variety of interpretations: *monolingual*; *bilingual*; *multilingual*; *plurilingual*; *English as an Additional Language*; *immersion* (students with extensive exposure to the target language in school and beyond). In preliminary discussions with Cambridge ESOL, two areas have been identified where there might be a mutual interest :

1. Establishing the relationship between cognitive levels of understanding in particular domains, and levels of linguistic understanding, and whether this relationship varies between domains and between curriculum stages.
2. Benchmarking CIE qualifications in relation to levels of the Common European Framework (CEFR). On the surface, one should lead logically from the other. A suggested starting point might be the analysis of CIE candidate responses in terms of what they say about language competence.

5. Marking and awarding

In addition to the development of improved systems for data collection and management and the analysis and evaluation of test materials and candidate performance, marking and awarding processes afford a range of potential research investigations:

- exploring ways in which the future availability of item-level data can be used in the grading process;
- considering issues relating to the administration of tests within time zones;
- analysing the evidence base used for awarding in CIE qualifications and patterns of use;
- evaluating protocols for award processes, feeding into routine review and enhancement of awarding by CIE officers, chairs, etc;
- examining the role and format of Principal Examiner reports in grading/awarding;
- developing appropriate methods (and associated protocols and manuals) for better understanding of what is happening in marking in different CIE qualifications;
- investigating the possibility of establishing a control group of Centres for each CIE qualification to act as an aggregate of benchmark Centres;
- reporting of A* at A Level and AS and of reporting of UMS marks;
- assessing the validity of some statistical methods for detecting malpractice;
- piloting the use of a rank-ordering method to obtain judgemental grade boundaries in the awarding process using small entry CIE syllabuses.

6. Commissioned research

Ministries occasionally request the provision of more information about their relative performance internationally. There is, therefore, a need to identify what CIE can easily and reliably produce annually and through a format that is simple to use by ministry officials who are statistically naïve and that encourages good use of the data to inform policy and priorities.

The Hong Kong secondary education system is currently undergoing reform. It is proposed that all students will be expected to remain in school until the end of their sixth year of secondary education, when there will be a single baccalaureate-style examination: the Hong Kong Diploma of Secondary Education. Concomitant with changes to the curriculum will be changes to assessment. The Diploma will integrate several important changes including changes to the curriculum and to the subjects that candidates will take; the introduction of a component of school-based assessment for each subject; and, moving to a standards-referenced approach to reporting results. In the context of HKDSE, it is envisioned that future CIE research will address the issue of standards equating and details about moderation of the proposed question papers.

The provision of a range of high-quality examinations is undoubtedly a team effort involving an extensive array of operational, assessment and administrative personnel. It is important, therefore, that all key people are involved at the initial stages of any new research and provide the input necessary to ensure that CIE assessments end up being suitable for their intended purpose. For this reason, any information gathered from CIE

staff about proposed new research is of great importance and feeds directly into decisions about future programmes. Engaging other professional staff in research activities is thus instrumental in the sharing of professional expertise both within CIE and within the wider Cambridge Assessment organisation.

On a final note, a vital component in the research programme is the publication of research outcomes. The importance of disseminating findings from work already undertaken and, more importantly, the recommendations which result from that work cannot be understated. A number of papers in various journals and conference proceedings facilitate the sharing of CIE research and international practice.

Research News

Conferences and seminars

House of Commons Research Seminar

The fourth House of Commons Research Seminar, chaired by Barry Sheerman MP, Chair of the Children, Schools and Families Select Committee, took place on July 1st 2008. The seminar, which was on the topic of what makes government initiatives succeed or fail, was attended by 60 key senior education professionals and MPs, generating a lively debate. Speakers included Kathy Sylva, Sue Burroughs Lange and Philip Davies.

They each gave their different perspectives on what it is that makes Government initiatives succeed and take root in mainstream practice, how the best cutting edge research coming out of institutions can be adopted by policy-makers and why sometimes ideas that appear to be beneficial when seen from a research perspective are not taken up by Government.

Professor Kathy Sylva talked about models for how researchers and policy makers can work effectively together. She used the Effective Provision of Pre-School Education Project, commissioned in 1996 – and still ongoing – as a case study.

Dr Sue Burroughs Lange of the Institute of Education outlined her experiences in trying to encourage the uptake of the Reading Recovery programme.

Philip Davies of the American Institutes for Research, who served in the Strategy Unit at the Cabinet Office, gave a presentation based on his experiences of evidence based policy making.

European Association for Research on Learning and Instruction (EARLI)/Northunbria Assessment Conference

Beth Black attended the Fourth Biennial Joint EARLI / Northumbira Assessment conference in Berlin in August and presented research on using an adapted rank ordering method to investigate January versus June awarding standards.

British Educational Research Association Conference (BERA)

In September eleven researchers from the Research Division presented papers at the annual BERA conference which was held at Heriot-Watt University, Edinburgh.

European Conference on Educational Research (ECER)

Martin Johnson attended the ECER conference at the University of Gothenburg in Sweden in September and presented a paper entitled: *A case of positive washback: an exploration of pre-release examinations on geography class room practice*.

International Association for Educational Assessment (IAEA)

The 34th IAEA Annual Conference took place from 7th–12th September at Robinson College, University of Cambridge (see page 2). The conference is a major event in assessment, bringing together leading assessment and education experts and providers of examinations from across the world.

Researchers from Assessment Research and Development attended the conference and presented papers covering a wide range of themes. See http://iaea2008.cambridgeassessment.org.uk for further details of the papers and presentations.

Association for Educational Assessment – Europe (AEA-Europe)

In November Sylvia Green and Tim Oates attended the 9th AEA-Europe conference in Hisar, Bulgaria. The theme of the conferences was: *Achieving quality in assessment: validity and standards*. Sylvia Green presented a paper on *Aspects of Writing: Beyond an atomistic approach to evaluate qualities of features of writing*.

Forthcoming conference

The 2009 Cambridge Assessment Conference will take place on Monday 19th October 2009 at Robinson College, Cambridge. Further details will follow in the next issue of *Research Matters*, or contact the Cambridge Assessment Network at: thenetwork@cambridgeassessment.org.uk.

Publications

The following articles have been published since Issue 6 of *Research Matters*:

Black, B. and Bramley, T. (2008). Investigating a judgemental rank-ordering method for maintaining standards in UK examination. *Research Papers in Education: Policy and Practice*, **23**, 3, 357–373.

Crisp, V. (2008). Exploring the nature of examiner thinking during the process of examination marking. *Cambridge Journal of Education*, **38**, 2, 247–264.

Greatorex, J. and Bell, J.F. (2008). What makes AS marking reliable? An experiment with some aspects of the standardisation process. *Research Papers in Education: Policy and Practice*, **23**, 3, 333–355.

Suto, W.M.I. and Nádas, R. (2008). An exploration of self-confidence and insight into marking accuracy among GCSE maths and physics markers. *Magyar Pedagógia*, July–August.

Suto, W.M.I and Greatorex, J. (2008). What goes through an examiner's mind? Using verbal protocols to gain insights into the GCSE marking process. *British Educational Research Journal*, **34**, 2, 213–233.