Issue 1 September 2005

Research Matters

CAMBRIDGE ASSESSMENT

University of Cambridge
Local Examinations Syndicate

Research Matters : 1

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 copies of articles are available on our website
www.cambridgeassessment.org.uk/research

Foreword

Welcome to the first issue of *Research Matters*, a biannual publication from Cambridge Assessment. The aim of this publication is to share assessment research in a range of fields with colleagues within Cambridge Assessment and in the wider assessment community and to comment on prominent research issues. It includes short summaries as well as more lengthy articles and papers reporting on current and completed research. Contributions are from Assessment Directorate research staff working across many areas of assessment and examinations. I hope that you will find *Research Matters* interesting and informative.

Ron McLone *Director-General of Assessment*

Editorial

In this issue we report on a wide range of research topics from standards over time to auto-marking of short textual responses. In the opening article Alf Massey describes a major, influential research study on the comparability of national test standards over time. This was a three year study commissioned by QCA and was carried out from 1999 to 2002. The article outlines the methodology and main findings, and considers the implications and impact of a study that is still giving rise to debate.

The theme of standards, reliability and validity are continued in Tom Bramley's summary of his article 'Accessibility, easiness and standards', published in *Educational Research*. He considers the challenges of setting cut-scores as part of the test development process and discusses the meaning of 'standards'. The summary of his second article, 'A rank-ordering method for equating tests by expert judgement', published in the *Journal of Applied Measurement*, investigates the use of judgements in scale construction, in particular an extension of the Thurstone paired comparison method to rankings of more than two objects.

In her article on grade descriptors Jackie Greatorex continues the theme of expert judgement in a different context. She outlines four studies about developing grade descriptors and using them for teaching and learning. This is followed by Victoria Crisp's and Ezekiel Sweiry's article that considers issues of validity and question difficulty with regard to the use of illustrations in examinations. This work formed part of an extensive research programme that focussed on factors unintentionally affecting the difficulty of examination questions.

Turning to John Bell's article, 'Gold standards and silver bullets', we have a discussion on a topical issue, the problem of the assessment and selection of high attainers. In his article John considers a number of options to overcome the problem as well as their potential consequences.

The final article, by Jana Sukkarieh and Stephen Pulman of Oxford University and Nicholas Raikes of Cambridge Assessment, introduces an UCLES-funded project that investigated automatic marking of short answer textual responses. The results of a thorough evaluation will be reported in a future issue of *Research Matters*.

We finish with 'Research News' which includes conferencing information and details of the Cambridge Assessment Conference in October.

Sylvia Green *Head of Research Programmes Unit*

Comparability of national tests over time: a project and its impact

Alf Massey Head of Evaluation and Validation Unit

The Comparability Over Time (CoT) Project was commissioned by the Qualifications and Curriculum Authority (QCA) and the research team, based at UCLES, began work in April 1999. The project investigated the stability of national test standards at all key stages and in all subjects and a final report (Massey, Green, Dexter & Hamnett, 2003) was made public by the QCA in December 2003. Since then the project's findings have attracted (and continue to attract) much comment. National test standards are of considerable public interest, not least because of the political prominence these tests have been accorded, including government claims that the huge improvements in results since tests were introduced in the mid-1990s stem from the plethora of recent educational policy initiatives.

Why comparability over time has proved elusive

The search for comparability over time in large scale assessments in some ways resembles that for the Holy Grail: it has a somewhat mythological quality; it is difficult to find and even many of those who should know better are uncertain what it might look like. Comparability over time is thus famously difficult to investigate (Goldstein, 1983; Newton, 1997) and even the language we use to talk about educational standards is often ambiguous (Massey, 1994).

As is the case with UK public examinations, new versions of England's national tests are set annually. Inevitably difficulty will vary somewhat from year to year; posing an annual standard setting conundrum. Each year we must decide which mark merits the award of each level (i.e. set level thresholds) so as to 'maintain standards'. Yet even in this context, where whole national cohorts take successive versions of a test, it is difficult – probably impossible – to provide an absolutely sound methodology to guarantee equivalence over time. Decisions must disentangle the effects of changes in the quality of teaching and learning (themselves affected by national and local policy initiatives and resource issues etc.) from the effects of variations in things like:

- the curriculum itself, both formal and informal (i.e. how teachers choose to implement it);
- test content and questions and, hence, 'absolute' difficulty;
- the calibre of samples taking given test forms for trial or equating purposes;
- the format of assessment used – including deliberate enhancements and changes in style;
- wider aspects of culture and social expectations that influence children's responses.

Different methodologies

Judgemental comparison of the equivalence or level of demand of the question papers themselves is endlessly fascinating, but research teaches us very clearly that teachers or other experts cannot predict the likelihood of students answering correctly accurately enough to allow test thresholds to be set, safely, on this basis. Judgemental comparisons involving pupils' work are more feasible and have the advantage of enabling comparisons across diverse assessments and contexts, but they are rather blunt instruments, probably only able to detect gross changes. This limits reliance on them in standard setting (Cresswell, 2000). But such judgements remain of interest. They are often valued by decision takers and other interested parties and there are systematic approaches in development which show some promise (Bramley, 2005a). Christie and Forrest (1981) used expert judgements to compare GCE AL standards over a ten year period and their work showed how much the criteria implicit in examinations change and hence make comparisons more difficult. Analytic comparisons (e.g. Massey & Elliott, 1996) are rare, partly for lack of substantial archives of scripts from the past, but have something to offer in investigating the variations in the ways in which achievement is demonstrated over longer periods of time.

Indirect comparisons can provide interesting data, using reference measures (which remain the same from year to year and may take the form of prior achievement or other 'existing' measures, such as scores on standardised tests) taken by children in successive cohorts to estimate their relative ability/achievement. Do children with similar reference test scores fare equally well on each year's test? The CoT Project made considerable use of this approach, gathering extensive data thanks to the co-operation of Local Education Authorities (LEAs). But the effects of curricular change on what is learned and assessed are considerable, especially over intervals of more than a few years, and those preferring different outcomes can disagree about how such factors might affect test performance. Argument about the varying relevance of the common measure to assessments in different years often makes conclusions based on this approach indicative rather than conclusive (Newbould and Massey, 1979).

Direct experimental comparisons seem an obvious method, but have not previously been used on a large scale in the context of Britain's curriculum driven performance assessment systems because of both high costs and the rate of curricular change, which quickly makes 'past' tests out of date for today's children and hence invalidates comparisons. It is worth noting that it was not until 1996, the base year for our study, that the national test system approached stability. Comparisons like ours could not have been commissioned earlier and QCA should take credit for their promptness in initiating this work when they did. Quasi-experiments, equating next year's test forms to those taken operationally by today's children, have been part of national tests' standard setting

process since their inception. But these are bedevilled by pupils' varying motivation levels when taking real and trial tests. Performance appears to fluctuate unpredictably – perhaps according to variations in school culture which help determine how seriously pupils take non-operational tests. The CoT Project was fortunate to enjoy the resources and opportunity to use medium-term direct experimental comparisons as the major basis for comparisons, thus avoiding such problems.

The Comparability Over Time Project

The project's work involved two main strands of quantitative research, supported by two qualitative approaches:

Medium-term direct experimental comparisons

Randomly assigned groups of children (in Northern Ireland – NI) took either 1996 or later versions of national tests in all subjects at all key stages, in the course of three (annual) phases of experimental testing between 1999 and 2001, as shown below.

- KS1 Reading Comprehension 1996 v 1999
- KS1 Mathematics 1996 v 2000
- KS2 English 1996 v 1999 and 1996 v 2000
- KS2 Mathematics 1996 v 1999
- KS2 Science 1996 v 2001
- KS3 English 1996 v 2001
- KS3 Mathematics 1996 v 2000
- KS3 Science 1996 v 2001

These intervals were short enough that the tests being compared were not so dissimilar that comparisons were invalidated and long enough for detectable effects to have appeared.

Children in NI were used so that they had not seen either the current or previous versions of these tests beforehand. In all 11,762 children from 184 schools were involved – providing large enough groups (circa 1,000 for each experimental comparison – well distributed across the range of ability) for sufficiently powerful statistical comparisons. Detailed desk research, fieldwork and/or questionnaires for teachers were used to investigate the validity of our conducting research with children following the slightly different curriculum in NI; enabling some (largely minor) issues to be identified, ready to be considered alongside the empirical data when reaching conclusions. Teachers from the schools involved considered these tests to be appropriate for their pupils and whilst they had not been prepared for them as directly as their English counterparts, any consequent reduction in overall performance was irrelevant to our purpose – that of comparing standards across versions rather than estimating how well the children themselves could perform.

For each subject at each key stage, two experimental groups were formed, to take the 1996 or the later version of the test respectively, using spiral quasi-random assignment. Alternate boys and girls on the school or class register were assigned to each form of the test, to minimise gender, school/teaching group and neighbourhood effects by distributing them evenly. However supplementary data concerning performance on relevant NI national assessments (or date of birth in the case of KS1 children for whom NI assessments were unavailable) were also collected, to help compare the equivalence of experimental groups and to use as a control variable if required.

Analyses then used the NI national assessment data (or date of birth) to check the equivalence of the experimental groups, before comparing the distributions of levels each group achieved to establish whether or not any differences were statistically significant. Wherever possible, features of the data available were explored to try to shed light on the origins of disparities observed between test forms.

In subjects/key stages where such variations were observed the data also enabled the project to equate the mark scales of the tests set in different years. If level thresholds set in the later year did not correspond to marks equated to the equivalent thresholds set in the earlier version, it suggests that variations in test results across the years have either under or over estimated the progress made by schools. A varied pattern of results was obtained, across key stages and subjects, as summarised in our conclusions below.

Evidence from LEA's standardised testing in schools

A few of the LEAs in England continue to use standardised tests extensively, often to help allocate resources. In effect these can provide found 'common reference test' data to help compare national test results by looking to see if children with the same standardised test scores from different years obtain equivalent national test results. We canvassed LEAs and gathered data available spanning all or parts of the period 1996–2000 to help cross-validate our experimental evidence.

Different LEAs use different tests and in effect provide a series of case studies. LEA 1 provided data from 29,896 children relating to KS1 English; 29,926 children relating to KS2 English; and 20,788 children relating to KS2 Mathematics – between 1996 and 2000. LEA 2 provided data for 22,985 children regarding KS2 English between 1996 and 1998. LEA 3's data was for 4,772 children concerning KS2 English between 1997 and 1998. LEA 4 provided data for 52,950 children concerning KS1 English between 1996 and 1999. LEA 5's data was for 17,963 children concerning KS2 English and for 17,971 children regarding KS2 Mathematics – between 1998 and 2000. LEA 6 provided data for 13,904 children concerning KS2 English and 15,747 children regarding KS2 Mathematics – between 1996 and 1998.

Evidence across LEAs suggested that (perhaps against the curricular odds, given that schools' attention must have been switching towards national tests as the latter's importance was increasingly recognised by teachers) standardised test scores had risen over the period investigated, in itself suggesting rising standards of teaching and learning.

Despite the very different methodologies and assumptions involved, relating the data on standardised testing to children's national test results provided convincing support for the conclusions reached via the experimental comparisons for the same key stages/subjects, regarding both the size and nature of effects observed. There were also considerable similarities between the data for the various subjects/key stages from different LEAs, so replication (both between methodologies and across these case studies) further bolsters confidence in the project's evidence.

Teachers' judgements about the quality of scripts

A small scale study involved judgemental comparisons (by teachers) of 'representative' 1996 and 1999 KS2 English scripts at key mark points. These teachers' judgements also supported the conclusions reached via experimental comparisons.

Children's perceptions of evolving features in national tests

Children's views about the tests they take are rarely sought. Our project interviewed small samples of children regarding every subject at each key stage (*n* ranging from 12 to 24 in each case), using selected paired

comparisons between materials from 1996 and the more recent test materials. A modified version of Kelly's repertory grid questioning technique was employed to help them verbalise their thoughts and identify salient features and even children as young as seven proved capable of doing so effectively.

Overall the materials sampled from more recent tests tended to be preferred – with some exceptions of course, as earlier versions themselves often contained attractive features. In short, children had appreciated the efforts which had been made to 'improve' the tests over the years by making them more attractive, more user friendly and more accessible.

Such changes will affect both motivation and performance. For instance those which make it easier for children to understand what is required or simplify the ways they respond seem likely to make it easier to demonstrate competence. But such issues raise interesting questions regarding test standards. For instance, should test thresholds be adjusted to compensate for greater user-friendliness? Or should developments which help children show what they can do be seen as a valid means of recognising performance – which should be reflected in improving results? Bramley (2005b) has recently considered the implications of changes in accessibility for measurement models used in the national test context.

Conclusions

- Experimental comparisons suggested that KS1 Reading Comprehension test standards at level 2c were similar in the 1996 and 1999 versions, but those at levels 2a and 2b may have been a mark or two more severe in 1999, so that gains in national test results may under-estimate progress for abler children. The methodologically independent relationships identified between test results and LEA standardised test data were consistent with this.
- Experimental evidence suggested that KS1 Mathematics test standards in 2000 were at least equivalent to those in the 1996 version, and here too levels 2a and 2b were perhaps a mark higher than needed to equate to their equivalents in 1996. Level 3 appeared even more severe in the later version, by around three marks. LEA standardised test data were again consistent with this conclusion and the implication is that improvements posted in national results over this period are in general merited, and may indeed slightly under-estimate progress by abler children.
- Experimental evidence suggested that there were disparities in KS2 English test standards between the 1996 and 2000 versions. It would have been necessary to increase Level 4 and 5 thresholds in the 2000 version by five and seven marks respectively to equate them to 1996 standards. Differences of this order might account for about half the national gains in test results over the period. Given that the writing element in the tests remained almost entirely stable, it can be deduced that the difference was attributable to the Reading element. The project replicated comparisons with the 1996 version in 1999 and 2000 in great detail, which suggested convincingly that the experimental methodology was robust. Relevant data from LEA standardised testing programmes supported the nature and size of the effects observed experimentally here too. In a third independent methodology, albeit in a small-scale qualitative comparison, judgements by teachers asked to compare samples of 1996 and 1999 scripts representing key mark points also concurred with the larger-scale empirical data.
- Experimental data suggested that KS2 Mathematics test standards were similar in the 1996 and 1999 versions, despite the potential for disturbance brought about by the introduction of a mental arithmetic element during this period. Relationships between LEA standardised test data and national test results in successive cohorts supported this too.
- Experimental comparisons indicated that to equate to standards in the 1996 version, thresholds in the 2001 KS2 Science test at levels 2, 3 and 4 would have needed to have been set somewhat higher – by perhaps two marks at levels 2 and 3 and by four marks at level 4. However, 2001's level 5 threshold appeared in line with the standard set in 1996. But changes of this order would account for only a small proportion of the very large gains in KS2 national test results in this subject over the period concerned.
- Because it was impossible to use paper 2 (the Shakespeare element) in experimental comparisons in KS3 English, the methodology had to be adapted here and comparisons were based on predicted test levels generated from experimental comparisons involving only paper 1. Additionally, a small scale re-marking exercise investigated the possibility of 'expectation creep' in contemporary markers' judgements about pupils' writing, compared with those made by markers in 1996. The latter suggested that more demanding marking of writing may have offset slightly lenient thresholds in the reading element detected in the experimental comparisons, leading to the conclusion that no differences could be detected in overall test standards between the 1996 and 2001 versions.
- KS3 Mathematics tests involve a series of attainment-related Tiers, targeted at levels 3–5, 4–6, 5–7 and 6–8 respectively. The experimental comparisons suggested that KS3 Mathematics test standards in the 1996 version appeared more severe than the 2000 version, especially in the lower tiers. Those taking the 2000 version of Tier 3–5 achieved results about half a level better than those taking the 1996 version; in Tiers 4–6 and 5–7 results were about one quarter of a level better on the 2000 version; and in Tier 6–8 achievement was only about one tenth of a level better on the more recent version. Investigation suggested that perhaps half of the effects observed were attributable to the introduction of the mental arithmetic element mid-way through this period. The size of the overall effect would account for a significant proportion of the improvement in test results nationally over this period.
- Experimental comparisons suggested that KS3 Science test standards were similar in the 1996 and 2001 versions.
- The varied findings across key stages and subjects suggests that there has been no conspiracy to manipulate test results, otherwise all tests at all key stages would appear more lenient. It should be recognised that the task of those responsible for setting standards has been daunting given the pace of change as the new curriculum and testing regime has been introduced, refined and improved. Only now is the system settling.
- Investigations of children's perceptions of national tests via structured interviewing, in all subjects and key stages, showed them able to identify, and appreciate, salient features of the tests which have changed as national testing has evolved. Such developments largely aim to make the tests more accessible, more interesting and motivating and user-friendly; although the impact of such changes on performance is hard to gauge. The project's final report considers

various issues concerning such enhancements. It underlines the need to keep the effects of the renewal of curricular and assessment regimes in mind and to manage their impact so that they do not threaten the validity of assessments used for monitoring standards.

- The project team drew upon the project's evidence, and other experience, to make various policy recommendations designed to help maintain test standards. These include:
  - an integrated medium term cyclical approach to the management of curriculum renewal and national test development;
  - a strategy for test equating which replaces a year on year focus by a stepwise approach involving equivalence between successive tests and a 'stable' baseline measure, within each cycle of curriculum/assessment regime renewal, before moving – cautiously – to a new baseline when the cycle of curricular change necessitates;
  - a logical basis for prioritising evidence available when setting test thresholds: which pays due regard to national sample data and the inherent status quo, whilst simultaneously strengthening the search for sound (and transparent) evidence which should be seen as a pre-requisite for shifts in the pattern of results;
  - the longer term suggestion that teacher assessments (which the project's standardised test data suggested were relatively stable overall) might contribute to or determine national assessments for individual children, with (less intrusive and less costly) tests being used to monitor the system and direct any moderation of differences between schools and/or teachers.
- But the most important conclusion which can be reached from the project's work is tangential to our original brief. Taken alongside national test results over the period, the experimental evidence from all subjects and key stages indicated that there has been substantial real improvement in children's achievement. This was cross-validated by analyses of the LEA standardised test data made available to the project, wherever such data were relevant. The standardised test evidence, provided by several different sources, suggests that KS1 standardised reading scale scores improved by about 25% of a standard deviation, whilst in KS2 standardised reading scale scores improved by 10% to 16% of a standard deviation and standardised mathematics test scores improved by around 25% of a standard deviation. There have been significant gains in achievement in all subjects at all key stages, even those where our evidence suggests that national test results may be exaggerating their extent. Given that an extensive body of previous research demonstrates that system-wide improvements in achievement are generally small and hard won (Brooks, Foxman & Gorman, 1995), this should be seen as cause for congratulation to all concerned.

Impact

This project is likely to have a significant influence on research in this vein because of its methodological innovation. Not only is it the first large scale use of direct experimental comparisons, demonstrating their effectiveness over medium term time intervals of 3–5 years, but it is the first to collect allied analytic/empirical data to check on potential influences of curricular variations for the sample tested and any potential interactions with items contained in different test forms. This enabled systematic evaluation of the influence of such factors on the validity of the results. The project is also notable for the use of both replication and alternative methodologies to cross-validate comparisons. Confidence in research outcomes is much enhanced by replication, and when different approaches based on different data, assumptions, and definitions of equivalence, all point to the same conclusions – as was the case for this project, we can regard the findings as robust. The project's qualitative strand investigating pupils' capacity to identify and value enhancements to assessment instruments, and discussing how such issues relate to comparisons and the maintenance of standards in educational assessments, is also novel. More careful management of the introduction of enhancing features where comparisons over time are seen as a key use for assessments is likely to become a serious issue as awareness of their importance grows.

The project's substantive findings naturally attracted media attention. It made headlines even before the report's publication (e.g. 'Test result bombshell kept under wraps' – *Times Educational Supplement (TES)* 17.10.03) and attracted considerable news and editorial coverage on publication (e.g. 'Doubt cast on primary pupils progress' – *Guardian* 18.12.03).

It has continued to surface in news coverage from time to time since then and is now – two years on, in June 2005 – at the centre of a spat between the Statistics Commission (an official watchdog over the use of statistics by public bodies) and the Department for Education and Science (DfES). Professor Peter Tymms wrote to the Commission enclosing a paper (Tymms, 2004) – which extensively cites our conclusions – as evidence that national test statistics are unsuitable to represent trends in educational standards. The Commission's report (Statistics Commission, 2005) upheld key aspects of the complaint and suggested that official statements should indicate that improvements in KS2 test scores between 1995 and 2000 are in part due to factors other than rising achievement. (e.g.'QCA admits to 'illusory' primary test improvements' – *TES* 06.05.05). The QCA's evidence to the Statistics Commission had fulsomely supported the CoT Project's work and conclusions and, hence, the Commission's stance. However, the DfES objected to any restraint on the use of test results as evidence of the success of government policy and asked the Commission to think again: an approach the Commission has since largely rebuffed (e.g.'Commission stands by primary figures' – *TES* 03.06.05). So it seems unlikely that the matter has been laid to rest. Does it matter? Yes it does. Probity and confidence may be called into question if government appears to pick and choose between the convenient and inconvenient findings of well conducted research when assigning the credit and blame for past events. What of the implications for evidence-based policy making? For instance, might some of this project's conclusions have implications for the test targets which schools have been set for future years? Inappropriate targets are unlikely to serve their intended purpose. Policy making cannot ignore the facts.

Notwithstanding the brouhaha described above, the project's conclusions were in fact far from condemnatory about the management of the national testing system, which was in its infancy in the period concerned. Whilst problems were detected in some assessments, others successfully achieved the difficult task of maintaining test standards over a period of rapid change. The QCA and the (recently constituted) National Assessment Agency (NAA) and their contracted test development agencies have considerable technical expertise at their disposal and their good intentions are not disputed. But the project

(aided by the media coverage received – partial and inaccurate as it was at times) may well have a salutary effect on the thinking of those at the highest levels involved in policy making on assessment matters. Examining Bodies have long been aware that setting standards is a difficult and complex process; partly as a result of research into comparability issues. Putting national assessments under the spotlight will make the politicians and professionals managing them acutely conscious that the concept of error of measurement has real as well as theoretical aspects.

The Project suggested some quite fundamental improvements to arrangements for national tests. Some were implemented even before formal publication of the final report and it is understood that, partly in consequence of potential risks to the maintenance of standards having been highlighted by the CoT project, a wide-ranging review of the relevant features of the key stage test system is under way. Irrespective of whether our suggestions or alternative solutions are adopted, the project has served a valuable purpose in making policy makers aware of the need to treat the conceptual and technical aspects of educational standards with greater respect.

References

Bramley, T. (2005a). 'A rank-ordering method for equating tests by expert judgement', *Journal of Applied Measurement*, **6**, 2, 202–223.

Bramley, T. (2005b). 'Accessibility, easiness and standards', *Educational Research*, **47**, 2, 251–261.

Brooks, G., Foxman, D. & Gorman, T. (1995). *Standards in literacy and numeracy,* New Series 7. London: National Commission on Education Briefing.

Christie, T. and Forrest, G. (1981). *Standards at GCE A-level: 1963 and 1973*. London: Schools Council Publications/Macmillan Education.

Cresswell, M. (2000). 'The role of public examinations in defining and monitoring standards', *Proceedings of the British Academy*, **102**, 69–120.

Goldstein, H. (1983). 'Measuring changes in education over time: problems and possibilities', *Journal of Educational Measurement*, **20**, 4, 369–377.

Massey, A. (1994). 'Standards are slippery', *British Journal of Curriculum and Assessment*, **5**, 37–38.

Massey, A. and Elliott, G. (1996). *Aspects of writing in 16+ English examinations between 1980 and 1994*, Occasional Research Paper 1. Cambridge: UCLES.

Massey, A., Green, S., Dexter, T., and Hamnett , L. (2003). *Comparability of national tests over time: key stage test standards between 1996 and 2001*, final report to the QCA of the Comparability Over Time Project. Available from http://www.qca.org.uk/news/6301.html.

Newbould, C. and Massey, A. (1979). *Comparability using a common element*, Occasional Publication 7. Cambridge: TDRU .

Newton, P. (1997). 'Examining standards over time', *Research Papers in Education,* **12**, 3, 227–248.

Statistics Commission (2005). *Measuring Standards in English Primary Schools, Report No 23*, February 2005. London: Statistics Commission.

Tymms, P. (2004). 'Are standards rising in English primary schools?', *British Educational Research Journal*, **30**, 4, 479–493.

Accessibility, easiness and standards

Tom Bramley Principal Research Officer, Evaluation & Validation Unit

The following is a summary of a research article published in summer 2005 which was prompted by my experience of working in the National Curriculum test development group, formerly part of UCLES' Research & Evaluation Division, now part of OCR. One major task in the test development process is to carry out statistical and judgemental exercises which can provide evidence about where to set the cut-scores on the test. These cut-scores (level threshold boundaries, equivalent to grade boundaries on GCSEs and A-levels) are supposed to be at the same standard each year. Of course, tests can vary in difficulty and the cut-scores might not be at the same point on the raw mark scale each year, in order to allow for fluctuations in difficulty of the test from year to year (as with GCSEs and A-levels).

However, discussions about whether one year's test is easier or more difficult than the previous year's test can often get bogged down when the spectre of 'accessibility' raises its head. Is a 'more accessible' test the same as an 'easier' test? Are there any implications for where the cut-scores should be set if a test is deemed to be more accessible, as opposed to more easy? Is there any way to identify questions which are 'inaccessible'?

The main purpose of the article was to use a psychometric approach to attempt to answer these questions. The article begins by discussing the meaning of 'standards' and the ambiguity with which the term is used, particularly in media reporting of examination issues. The standard can be defined psychometrically as a point on the latent trait which is assumed to underlie or cause the responses to the test questions. The informal definition of statistical equating – that if standards have been correctly applied to two tests then it should be a matter of indifference to candidates whether they take test A or test B in terms of which level they obtain – is used as a starting point for discussing the issues raised by accessibility.

Three prototype arguments in favour of not raising the cut-scores by as many marks as the statistics might suggest when a test is deemed to be more 'accessible' were used to illustrate the discussion:

> *The paper is more accessible, but the amount of science hasn't changed.*
>
> *We've removed some of the hurdles which prevented the pupils from showing us what they can do.*
>
> *The pupils will be less 'turned off' by the paper and so we'd expect performance to improve.*

I have heard variations of these arguments on many occasions in the course of my time at UCLES – you might wish to pause here to consider whether you think they are valid.

In the article I argued that all of these arguments ignore the basic idea behind statistical equating that it should not matter to the pupil whether they take the more accessible or less accessible test. From a psychometric perspective, these arguments are all saying that the new accessible test is not measuring quite the same trait or construct as the old, less accessible test. To the extent that this is true, then strictly speaking it is not possible to set cut-scores on the two tests which have the same meaning.

However, it is at the level of the item/question that the actual changes in accessibility occur. The second part of the article explores the use of Rasch misfit statistics to investigate accessibility issues within a single test (i.e. it simplifies the situation from the actual one which occurs in practice where we have two different tests). It turns out that both types of misfit (underfit, and overfit) have the potential to diagnose problems with accessibility. Substantial underfit (lack of discrimination) often indicates either a poorly worded or ambiguous question which has confused the more able pupils, or an incorrect or incomplete mark scheme. Substantial overfit (discrimination which is too good!) might signify a different, but highly correlated dimension. For example, a teaching effect could produce overfit. If only the most able pupils in a school are taught a particular topic then only they will be able to answer a question on that topic correctly. This would produce good discrimination, but this good discrimination is arguably invalid, because the question would be inaccessible to the lower ability pupils.

Of course, such psychometric indicators are only the starting point for qualitative research aimed at deducing patterns and rules for identifying particular types of question and response format where accessibility issues are likely to cause a measurement problem.

I concluded the article by asserting that treating the standard setting/maintaining issue as a measurement problem provides a rational basis for understanding accessibility. Increasing accessibility does in fact make the test easier and cut-scores should rise in order to maintain standards.

Further reading

Bramley, T. (2005). 'Accessibility, easiness and standards'. *Educational Research*, **47**, 2, 251–261. Available from http://www.tandf.co.uk

STANDARDS OVER TIME

A rank-ordering method for equating tests by expert judgement

Tom Bramley Principal Research Officer, Evaluation & Validation Unit

The following is a summary of a research article published in summer 2005.

This paper built on much research carried out at UCLES over the past ten years on the use of judgements in scale construction. The main technique which we had used was Thurstone's paired comparison method whereby two objects are compared in relation to a single trait. Repeated comparisons of different pairs of objects by several (or many) judges can allow a single scale to be created, with the objects located at different points according to how many comparisons they 'won' and the location on the scale of the objects with which they were compared. Thurstone's method has been used for comparability exercises (comparing scripts at the same grade boundary from the same subject at different exam boards), and for research into the perennial question of standards over time (comparing scripts at the same grade boundary in the same subject specification in different years).

As those who have been involved in these studies can testify, the Thurstone approach can be very time-consuming and tedious for the panel of judges involved, because of the number of judgements required to form a satisfactory scale. My idea was to attempt to speed up the process by asking judges to place a set of objects into a single rank order, rather than requiring many separate paired comparisons. A second variation on the Thurstone process as it had been used in the studies above was to involve the entire mark range, rather than focussing on a particular boundary. This was to allow the two mark scales to be compared at all points by plotting the mark on the script against the 'judged measure' (the outcome of the ranking procedure).

NAA (the National Assessment Agency, responsible for the National Curriculum tests) requires its test development agencies to carry out a judgemental standard-setting exercise using practising teachers in order to supplement the statistical procedures used to derive cut-scores on the current year's test. The agencies are given some flexibility in the methods they use, so we decided to try out the rank-ordering method in this context. Scripts from the Reading component of the Key Stage 3 English test in 2003 (live test scripts) and 2004 (final pre-test scripts) were used as the objects to be ranked. Approximately 40 scripts from each year in total were involved in the exercise, one on each mark covering the effective mark range of each test. All the question mark totals were 'cleaned' from the scripts so the judgements would be based on perceived quality rather than simply adding up the marks.

Each judge (from a panel of twelve) was given four packs of ten scripts. Each pack contained five scripts from 2003 and five from 2004. No two packs of scripts were identical, but there was a lot of overlap across judges and packs in order to create the linking necessary to form a single scale from their judgments. The judgemental task was simply to put the ten scripts in each pack in order from best to worst. Tied ranks were allowed, but strongly discouraged, and in the event there were only two

or three tied rankings in all 48 rank orders. The contents of each pack were systematically varied in terms of both the overall level and spread of scripts from each year. The judges were warned not to make any assumptions about the contents of their packs – it was possible (for example) for all five scripts from one year to be 'better' than all five scripts from the other.

The data were analysed with two statistical methods, both based on the Rasch model. The first method converted each ranking into a set of paired comparisons and proceeded to analyse them as usual. The second method treated each ranking as a separate Rasch Partial Credit item. When the resulting measures from the two methods were plotted against each other the points lay on a straight line, showing that the two methods were giving substantively the same result.

More interesting was the outcome of the exercise, obtained by plotting the mark on the script against the judged measure, and fitting separate best fit lines for each year, as shown in Figure 1.

Since the judged measures are all on the same scale, the two raw mark scales can be equated (perhaps a weaker term such as 'linked' is more appropriate): the marks corresponding to the same measure are deemed to be equivalent. The equivalent mark on the 2004 test to any mark on the 2003 test can be found either by reading off the graph, or by using the regression equations for the best fit lines. In fact, in this case the two best fit lines were approximately parallel, separated by a vertical distance of around three marks, leading to the conclusion that the 2004 Reading component was about three marks easier at all levels than the 2003 Reading component. This agreed well with the (completely independent) evidence from statistical equating of pre-test scores, which had suggested that the 2004 test was around two marks easier.

The article contains a lengthy discussion of the difference between standard setting and standard maintaining, arguing that the rank-ordering method is more appropriate than most other judgemental methods for standard maintaining, and that standard maintaining is more appropriate than standard setting in the national testing context.

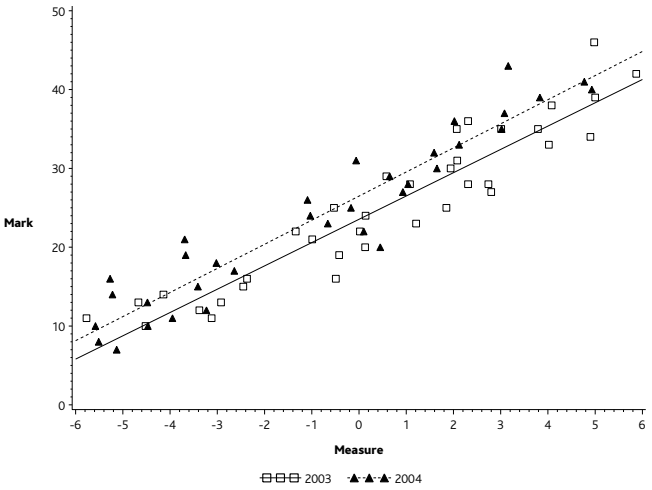


Figure 1: Plot of mark against measure for scripts from 2003 and 2004.

Since the paper was written, the method has been repeated successfully with the 2004 and 2005 Writing components of the KS3 English test, and is currently being investigated in a research study using scripts from two years of an A-level Psychology paper. There are also plans to investigate its suitability as an award meeting methodology.

Further reading

Bramley, T. (2005). 'A rank-ordering method for equating tests by expert judgement', *Journal of Applied Measurement* **6**, 2, 202–223. Available from http://www.jampress.org

DESCRIBING ACHIEVEMENT

A review of research about writing and using grade descriptors in GCSEs and A levels

Dr Jackie Greatorex Principal Research Officer, Research Programmes Unit

In this article I describe current Awarding practice and review some of the literature about writing and using grade descriptors (often also referred to as 'grade descriptions') for GCSEs and A-levels. Particular emphasis is given to the research that has used empirical evidence to write grade descriptors and the associated research methods.

> *Grade descriptors are descriptions of the qualities expected at different levels of a candidates' performance in an assessment* (Greatorex et al., 2001, 167).

The following are some extracts from grade descriptions for GCSE Biology:

Grade F: Candidates recall a limited range of information. For example, they state the main functions of organs of the human body and describe some defence mechanisms of the body (*OCR, 2000, 17*).

Grade C: Candidates describe how evidence is used to test predictions made from scientific theories, and how different people may have different views on some aspects of science (*OCR, 2000, 18*).

Grade A: Candidates use detailed scientific knowledge and understanding in a range of applications relating to scientific systems or phenomena. For example, they explain how temperature or water content is regulated in humans (*OCR, 2000, 18*).

Awarding

The Code of Practice (QCA, 2005) sets out the procedure by which grade boundaries should be determined. The Awarding Committee (senior examiners) must be provided with a variety of information to set the boundaries, including, where available, performance descriptions or grade descriptions. The Awarding Committee scrutinises scripts within a range of marks around the proposed key grade boundaries (e.g. A/B and E/U for A-level). They start at the top of the range of marks, scrutinising scripts on each mark in turn and agree on the lowest mark that is worthy of a higher grade. This is the upper limiting mark. Then they start at the bottom of the range, scrutinising scripts on each mark in turn and agree on the highest mark that is not worthy of the higher grade. The mark above this is the lower limiting mark. As a group they use their professional judgement to recommend a grade boundary within the range between the higher and lower limiting mark. The grade boundaries for grades which are not key grades are determined by taking the mark interval between key boundaries and dividing it equally between the grades. There are detailed procedures given in the Code of Practice explaining the rules to apply when the mark interval cannot be divided equally between the grades.

Writing grade descriptors

There have been attempts to write prescriptions of candidates' performances to be associated with different grades (grade related criteria) for GCSEs. For further information regarding the development of the grade related criteria and the associated limitations see Gipps (1990), Kingdon and Stobart (1988) and Cresswell (1987). The difference between grade descriptors and grade criteria is an important distinction to make. 'Grade criteria' are qualities a candidate's work *must* exhibit to be awarded a grade. 'Grade descriptors' are indicators which exemplify the qualities candidates are likely to exhibit if they achieve a particular grade. Normally grade descriptors refer to mid range performance within a grade rather than borderline performance. This article focuses on grade descriptors.

Massey (1982) attempted to use empirical evidence to describe the performance of candidates who achieved particular grades. He analysed marks at the question level, adopting a concept of group mastery as the aim was to describe the achievement of grade groups, not individuals. A grade group is all the candidates who were awarded a particular grade. If the mean mark achieved by a grade group was 75% of the total marks available for a question this was taken as an indication of group mastery. The arbitrary working value of 75% suggests that most of the candidates can answer a question correctly. There were two criteria for mastery:

- questions had to be mastered by all grade groups higher than the mastery level grade group and all groups below had to fail to reach the 75% criterion;
- questions had to discriminate statistically between the mastery level grade group and the group below (Massey, 1982).

The skills and knowledge required to answer questions where a given grade group defined the mastery level can be used to describe the grade group's competency. For the lower ability range the analysis did not indicate what candidates could master because on these tests the lower ability candidates generally got less than 75% on all questions. Massey recommended that other approaches would need to be used to capture this information. He argued that if other educators produced written descriptions of performance at the grades featured in his study, their descriptions would need to be reconciled with his results, or they would lack validity.

Pollitt and Murray (1996) argued that grade descriptors should match what assessors perceive in the performance they assess. Their method for writing grade descriptors was designed with this in mind. They asked assessors to compare pairs of candidates' work in the field of testing English as a second language. In each comparison they indicated which performance was better and which was weaker. The judgements of which performances were the better in each pair were statistically analysed to create a scale on which the performance of the candidates could be located. Immediately after making a judgement the assessors were interviewed using Kelly's Repertory Grid (Kelly, 1955) to describe the two performances. Kelly's Repertory Grid is a method of interviewing research participants in a systematic way to compare how objects – in this case candidates' work – are similar to and different from one another. It is a way of eliciting peoples' personal constructs. They found that particular characteristics of performance seem to be allied to different sections of the scale. The characteristics of low performance became increasingly less relevant at the higher performance end of the scale, while different characteristics exemplified higher performance and were evident only at the higher end of the scale. Pollitt and Murray argued that assessment judgements would be more accurate if grade descriptors referred only to the characteristics which are normally salient at each stage on the performance scale.

The following section refers to three articles about developing grade descriptors for A-levels and a fourth article about their use in teaching.

(1) Making the grade – Developing grade descriptors for Accounting using a discriminator model of performance, Greatorex, J., Johnson, C. and Frame, K. (2001), *Westminster Studies in Education*, **24**, 2, 167–181.

One of the purposes of the research was to establish whether the writing of grade descriptors for an Accounting A-level, examined in June 1998, was aided by the discriminator model of performance. The *discriminator model of performance* is a term we used to refer to Pollitt and Murray's argument that candidates exhibit distinctive qualities at different stages on a performance scale. Pollitt and Murray do not use this term themselves. The methods we used draw from both Massey, and Pollitt and Murray.

We took a random sample of candidates who achieved each component (question paper) grade A, B, C, D, E and O/N. The random sample for each grade is assumed to represent candidates' achievement at that grade on each question in the examination paper. A series of statistical criteria for mastery, similar to those used by Massey, were applied to these data. This stage of the research is called a *mastery levels analysis*. The outcome of the analysis is a list of examination questions which particular component grade groups have 'mastered'. On these questions candidates from adjacent grade groups are exhibiting different knowledge or skills.

The second stage of the research was to describe the qualitatively different knowledge and skills exhibited by candidates from different component grade groups. Two senior Accounting examiners were presented with two answers to a question from candidates in the grade group which mastered the question and one script from a candidate from the grade group below. These answers had been credited with the mean

number of marks for their grade group on that question. The answers were then compared using an adapted version of Kelly's Repertory Grid. The following extract is an example of which candidates' answers were compared:

> *two E grade scripts where the candidates had scored 4 marks on question 4a on component 3 and one grade O/N script where the candidate had scored 2 marks on question 4a.*
> (Greatorex et al., 2001, 175)

The examiners were interviewed and asked to describe how the answers to each discriminating question on the higher grade scripts were similar to one another and different from the answer on the lower grade script. The interviews were recorded. The procedure was repeated using a small number of scripts for each grade. This provided a record of the knowledge and skills which distinguished the performance of candidates who achieved particular grades from the performance of candidates awarded the grade below. From this record the researchers and senior examiners wrote grade descriptors.

Extracts from the grade descriptors are given below:

> ***Grade E:*** *Knowledge of what is required, understanding, selecting appropriate knowledge, some application, but can be an incomplete answer.*
>
> ***Grade B:*** *Can rise to the challenge of a novel situation, but with some evidence of technical and calculation errors .*
>
> ***Grade A:*** *Awareness of multi-faceted aspects of Accounting* (Greatorex et al., 2001, 176–177).

Writing grade descriptors grounded in empirical evidence is arguably an improvement upon methods which are based upon examiners' expectations alone. However, one problem of using empirical evidence in this way is that it is a post hoc method with grade descriptors based on what candidates achieved rather than on what the examinations were designed to assess.

The grade descriptors developed in this study were validated by using them in the Awarding Meeting for the Accounting A-level examined in June 1999. They were generally received positively because they proved to be appropriate for the 1999 scripts. This implies that the discriminator model of performance is a sound basis for a method of developing grade descriptors in this domain. The grade descriptors which were developed using the discriminator model of performance suggested that there are indeed different characteristics associated with each grade. However, in some cases the differences might relate to 'relative' performance. The following extract gives an example of relative performance:

> ***Grade A:*** *Clarity and brevity of expression in a focused answer* (Greatorex et al., 2001, 176).
>
> ***Grade B:*** *Focused answer* (Greatorex et al., 2001, 176).

(2) Making Accounting examiners' tacit knowledge more explicit: developing grade descriptors for an Accounting A-level, Greatorex, J. (2002), *Research Papers in Education*, **17**, 2, 211–226.

In the article referenced above I argued that the process of writing grade descriptors is a way of making senior Accounting examiners' tacit knowledge, or personal constructs about achievement at different grades, more explicit. Other sources of this information can be found in the specifications, question papers and mark schemes. Moreover, it is fair to share this knowledge so that teachers and candidates are aware of the qualities expected to be credited with a particular grade. The method described in this article differs from that used in the previous article in that the subject officer and researcher, rather than the examiners, collated the records from the interviews to write grade descriptors. It can be argued that the grade descriptors would be more valid if they were formulated from the interview records by the examiners themselves. The usefulness of the grade descriptors at Awarding was limited, since the style of the question papers had changed between data collection and validation. Nevertheless, they did provide a helpful reference point.

(3) Making the grade – How question choice and type affect the development of grade descriptors, Greatorex, (2001), *Educational Studies*, **27**, 4, 451–464.

In this study I aimed to develop grade descriptors for an A-level in Economics based on examination data and scripts from the summer of 1999. The Economics A-level had three examination papers: a multiple choice paper, a paper with a points based mark scheme and a third paper where the mark scheme was based around four generic level descriptors. Grade descriptors were not developed for the multiple choice paper but were developed for the other item types using the methods utilised in the two studies to write grade descriptors for Accounting described previously. In contrast to the studies in Accounting, the Economics grade descriptors were validated by using them in two Awarding Meetings in the next session, one was the same Economics A-level specification used for data gathering and the other was a different Economics A-level.

The Accounting papers had contained numerical questions as well as long and short written answer questions and the Economics questions were short answers and essay questions. I argued that the method of mastery levels analysis and Kelly's Repertory Grid technique can be used together for examinations with all of these types of questions and for subject domains which are orientated towards both numerical work and extended prose. One advantage of the grade descriptors developed in these studies is that they are written at the component level (i.e. the level at which the judgements are made in Awarding Meetings) which is different from the general practice of writing grade descriptors at the specification level.

Developing grade descriptors from empirical evidence using sound methods is good practice. It is also important that the grade descriptors are used appropriately and in the following section I refer to an article which addresses the issue of using grade descriptors in teaching.

(4) Can different teaching strategies or methods of preparing pupils lead to greater improvements from GCSE to A-level performance? Greatorex J. and Malacova E. (in press) *Research Papers in Education*.

In this study the aim was to investigate whether there was any relationship between relative progress from mean GCSE scores to A-level results on the one hand and teaching strategies or how teachers prepared pupils for assessments on the other. We sent a questionnaire to Chemistry teachers to survey the teaching strategies they used and how they prepared pupils for the different A2 units which are part of the A-level. The questionnaire responses were matched to A-level and GCSE results. A series of multilevel models were fitted to the data to identify any relationship between relative progress from mean GCSE scores to A-level results and the questionnaire responses. We found some activities which related to higher relative progress from mean GCSE to A-level unit marks. One of these was using grade descriptors to inform the teacher's

preparation of pupils for the synoptic unit. However, it was also found that there was no such effect for the other two A2 units, one of which was coursework.

The Chemistry grade descriptors had been developed using examiners' expectations alone rather than based on empirical evidence. Nevertheless, our findings showed that grade descriptors can be important and helpful to teachers and can enhance classroom practice.

Conclusions

It can be deduced that, when resources allow, it is good practice to write grade descriptors based on empirical evidence. It seems that grade descriptors for different domains and types of questions can be written using a combination of a mastery levels analysis and Kelly's Repertory Grid technique. The grade descriptors developed using these methods describe the distinctive characteristics of achievement at particular grades.

Despite the difficulties of effectively communicating the meaning of grade descriptors to examiners, teachers, candidates and other stakeholders, it is good practice to make efforts in this area.

There is little research about how grade descriptors are used, or could be used, in relation to teaching GCSEs or A-levels, or in preparing pupils for assessments and there is room for further research in this area.

References

Cresswell, M. J. (1987). 'Describing examination performance: grade criteria in public examinations', *Educational Studies*, **13**, 3, 247–265.

Gipps, C. (1990). *Assessment: a Teacher's Guide to the issues*. London: Hodder and Stoughton.

Greatorex, J. (2003). 'Developing and applying level descriptors', *Westminster Studies in Education*, **26**, 2, 125–133.

Greatorex, J. (2002). 'Making Accounting examiners' tacit knowledge more explicit: developing grade descriptors for an Accounting A-level', *Research Papers in Education*, **17**, 2, 211–226.

Greatorex, J. (2001). 'Making the grade – How question choice and type affect the development of grade descriptors', *Educational Studies*, **27**, 4, 451–464.

Greatorex, J., Johnson, C. and Frame, K. (2001). 'Making the grade – Developing grade descriptors for Accounting using a discriminator model of performance', *Westminster Studies in Education*, **24**, 2, 167–181.

Greatorex J. and Malacova E. (in press). 'Can different teaching strategies or methods of preparing pupils lead to greater improvements from GCSE to A-level performance?', *Research Papers in Education*.

Kelly, G. (1955). *The psychology of personal constructs*. New York: Norton. Reprinted by Routledge (London), 1991.

Kingdon, M. and Stobart, G. (1988). *GCSE Examined*. London: The Falmer Press.

Massey, A. J. (1982). 'Assessing 16+ Chemistry: The exposure-mastery gap', *Education in Chemistry*, September, 143–145.

Oxford Cambridge and RSA Examinations (2000). OCR GCSE in Biology 1980, www.ocr.org.uk

Pollitt, A., and Murray, N. L. (1996). 'What raters really pay attention to', in M. Milanovic, & N. Saville (Eds), *Studies in Language Testing: 3 Performance Testing, Cognition and Assessment: selected papers from the 15th Language Testing Research Colloquium*. Cambridge: Cambridge University Press.

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

ISSUES IN QUESTION WRITING

Can a picture ruin a thousand words? The effects of visual resources in examination questions

Victoria Crisp and Ezekiel Sweiry Research Officers, Research Programmes Unit

Introduction

Visual resources, such as pictures, diagrams and photographs, can sometimes influence students' understanding of an examination question and their responses (Fisher-Hoch, Hughes and Bramley, 1997). Visual resources are sometimes included to test students' abilities to interpret them, but they are more commonplace than this alone would warrant.

Research on the influences of graphics in instructional texts provides some relevant insights. Such research has often found illustrations to have a positive influence on learning and retention (Weidenmann, 1989; Ollerenshaw, Aidman, and Kidd, 1997). However, the main purpose of examination questions is to assess learning rather than teach. Graphics are thought to 'simplify the complex' and 'make the abstract more concrete' (Winn, 1989, p. 127). Graphics can also provide more information than can be explained in words (e.g. Stewart, Van Kirk and Rowell, 1979). These are justifiable reasons for including visual resources in examinations as they can reduce the length of questions and help students to access abstract concepts. In addition, illustrations are generally believed to have a motivational role in the context of instructional texts (Peeck, 1993) which could apply to examinations.

In their review of work in this area Levie and Lentz (1982) found that in about 15% of studies there were no significant effects of including illustrations. One possible explanation is that the quality and appropriateness of the graphic is important (see Peeck, 1987 for some evidence of this). Such failures have also been explained as either a result of students' learning styles (as Ollerenshaw, Aidman and Kidd, 1997 report) or due to students not processing graphics adequately (Weidenmann, 1989). The latter is thought to be a result of the apparent ease of processing an illustration, giving students the false impression

that they have fully understood an illustration when they have not (Weidenmann, 1989).

The main risks of including illustrations in examinations are that a graphic may lead to the formation of a mental representation of a question that does not match the meaning intended by the question setters, or that students may use a particular aspect of an illustration that was not intended to be important. When a student reads a question, a mental model is constructed as a response to the text (Pollitt and Ahmed, 1999). This representation is based on ideas that are already known to the reader (Johnson-Laird, 1981) and hence students' mental representations of the text (and any illustration) may not all be the same, perhaps emphasising certain aspects that are particularly salient to them. Most of this process is unconscious and automatic.

Visual resources are likely to play a large role in the development of the student's mental model of a question with more emphasis being placed on the ideas communicated by them than the ideas conveyed by the associated text. As Peeck (1987) states, "too much attention may be deployed to the illustrations themselves rather than to the accompanying text" (p. 118). There are a number of possible explanations for the apparent superiority of illustrations over text. Firstly, processing visual material may require less cognitive effort. According to Biedermann (1981) the general meaning of an image can usually be grasped in as little as 300 milliseconds. This may be because the elements of a visual resource can usually be processed simultaneously, whereas a text must be processed sequentially (Winn, 1987). Another perspective is that visual and textual materials may be processed in different cognitive systems. Paivio's (1975) theory of dual-coding explains the superiority of memory for images as a result of them being coded both as images and as their verbal labels whilst words are only encoded verbally, thus resulting in bias towards information gained from illustrations (Schnotz, 1993).

In general, placing information higher on a page will make it seem more important (Winn, 1987). However, there is also some evidence that visual resources are more likely to be 'read' and processed before accompanying text regardless of their relative positions (see Kennedy, 1974). It has been well documented that the first elements contained within a mental model will dominate and strongly influence subsequent elements (Gernsbacher, 1990). Hence the fact that visual resources are likely to be processed first means they will be likely to dominate the representation.

If visual resources do have a disproportionately large influence on the development of mental models, this has implications in examinations where students' ability to process material effectively is already compromised by test anxiety (Sarason, 1988). Students need to understand questions in the way intended in order to have a fair opportunity to display their knowledge and skills.

Method

525 students, aged 16 years, sat an experimental science test under examination conditions. The test included six questions involving graphical or layout elements. For most of the questions, two versions were constructed in order to investigate the effects of changes to visual resources on processing and responses. The questions were compiled into two versions of a test paper which were assigned to students at random. Twenty-seven pairs of students were interviewed after they had taken the test.

The predicted GCSE grades of the students were converted into scores from 8 points for a grade A* to 1 point for a grade G. The mean score for students attempting the test was 4.50 (N = 261, SD = 1.369) for version 1 of the test and 4.55 (N = 254, SD = 1.353) for version 2, suggesting that the two groups were very similar in ability.

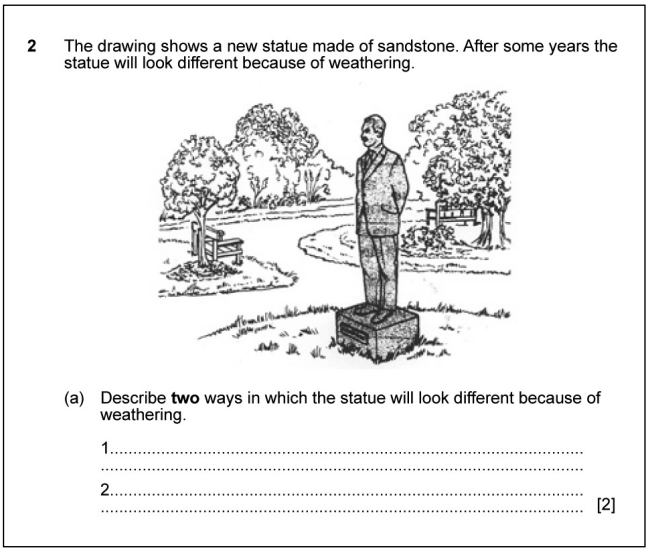
Results

For the purposes of brevity and interest four of the six questions will be discussed here.

Question 2 – Statue

This question was included to investigate the use students make of an illustration that is not essential. Answers such as 'discoloured', 'material worn away' and 'bits broken off' were credited. Responses referring to a named feature being altered were also accepted. Students did not necessarily need to use the graphic but it could help them to gain one of the marking points.

Students scored well on this question (85.5% of students scored both marks). There was a mixture of responses during interviews with regard to the illustration. Some said they found it useful, with slightly fewer stating that they did not find it of use or that it was unnecessary.


2 The drawing shows a new statue made of sandstone. After some years the statue will look different because of weathering.

(a) Describe **two** ways in which the statue will look different because of weathering.

1.....

.....

2.....

..... [2]

Positive comments included feelings of reassurance at having an illustration, and mention of how they could see that the statue had detailed features that might be lost. For example, one said, *it gave you an idea of what to pick out... Like you would lose a lot of detail on the person*. Some students held quite strong views that the graphic was unnecessary. For example, one claimed that *You don't even need the picture. I mean if you say a statue it could be of anything, it works the same way.*

There is also a possibility that the use of an illustration may affect whether students use scientific reasoning. One student commented, *I think if I didn't have the diagram there I would probably use more science instead of just saying its features won't be so defined.*

Another interesting comment supported the view that the inclusion of an illustration could reduce the amount of attention paid to the text: *it might lead you to look at the picture instead of the text so the answer might be wrong because you haven't read the text properly.*

Question 5 – Children's meal

Version 2 of question 5 included an unrealistically large sized portion of chicken nuggets in the visual resource to investigate whether the salience

of this might dominate students' thinking and lead to answers about overeating. Version 1 acted as a control.

No responses about overeating occurred with either version of the question and the marks scored on each version were very similar. Most of the interviewees who had attempted version 2 had not noticed the large portion size. In addition, most comments suggested that students were viewing the graphics as generic illustrations of the food types rather than as the actual meal. After the portion size had been mentioned by the interviewer, one student said, *It doesn't really matter though, it just shows, illustrates what it is but it doesn't really matter how much there is*. Several students made comments along the lines of, *I didn't really look at the pictures, I went straight to the ingredients*.

Additionally, the students who did notice the portion size did not use this in answering, perhaps not expecting an answer relating to quantity of food to be relevant in science. One student commented that *You're thinking about it as in science so it would be like the content in it not the amount*.

Question 5 : Version 1

5 Use the information below to help you answer the following question.

Children's meal		
Chicken nuggets		Chicken, wheat flour, maize flour, hydrogenated vegetable oil, salt, modified starch. Raising agents, mono calcium phosphate, sodium bicarbonate, sodium aluminium phosphate. Starch, spices, whey powder, pepper, dextrose, vegetanbe oil. Acidity regulator, calcium lactate. Emulsifiers, phosphate salt.
French Fries		Potatoes (cooked in our own vegetable oil) dextrose, Salt
Milkshake		Milk, skimmed milk, cream, sugar, skimmed milk powder, glucose. Stabiliser, guar gum, sodium polyphosphate, carrageenan and carboxymethylcellulose. Vanilla flavour.

Give **two** reasons why it would not be advisable for a child to eat this meal every day.

1.....
.....
2.....
..... [2]

Question 5 : Version 2

5 Use the information below to help you answer the following question.

Children's meal		
Chicken nuggets		Chicken, wheat flour, maize flour, hydrogenated vegetable oil, salt, modified starch. Raising agents, mono calcium phosphate, sodium bicarbonate, sodium aluminium phosphate. Starch, spices, whey powder, pepper, dextrose, vegetanbe oil. Acidity regulator, calcium lactate. Emulsifiers, phosphate salt.
French Fries		Potatoes (cooked in our own vegetable oil) dextrose, Salt
Milkshake		Milk, skimmed milk, cream, sugar, skimmed milk powder, glucose. Stabiliser, guar gum, sodium polyphosphate, carrageenan and carboxymethylcellulose. Vanilla flavour.

Give **two** reasons why it would not be advisable for a child to eat this meal every day.

1.....
.....
2.....
..... [2]

One student made an interesting comment in favour of including graphics in examination questions: *the use of pictures isn't particularly useful in trying to answer the question, but it's quite daunting on the day if all you've got is text and you've just got to read it, so maybe a picture would calm your nerves*.

Students, in the case of this question, seemed to know that they should not place emphasis on the information in the illustration.

Question 6 – Products

The word 'products' has a specific meaning in chemistry but a more familiar meaning might be that of 'household products'. Both are likely to be triggered in students' minds with question 6, although those with sufficient subject knowledge are likely to be able to suppress the irrelevant idea. The aim of including this question was to investigate whether the use of an illustration of some household products (in version 1) could affect the interpretation of the word 'products'. This would, of course, be undesirable in an examination but could occur by accident if care was not taken when choosing a visual resource. Version 2 acted as a more neutral control question.

Question 6 : Version 1

6 After eating a meal, your mouth becomes very acidic. This acid can damage your teeth.

[Part (a) omitted]

Brushing your teeth with toothpaste will neutralise the acid. This will protect your teeth from damage.

[Part (b) omitted]

Some brands of toothpaste contain sodium carbonate.

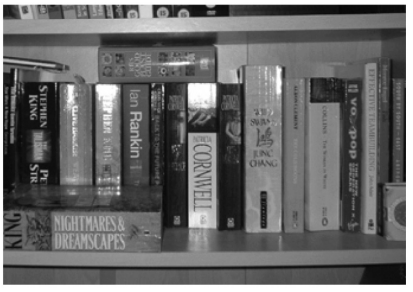
(c) Three products are made when sodium carbonate reacts with hydrochloric acid.

What are they?

1.....
2.....
3..... [3]

Question 6 : Version 2

6 The paper in modern books contains slight traces of acid. The acid in the paper can make it slowly decay.



[Part (a) omitted]

One method of neutralising the acid in books is to use sodium carbonate.

[Part (b) omitted]

(c) Three products are made when sodium carbonate reacts with hydrochloric acid.

What are they?

1.....
2.....
3..... [3]

In version 1, 9% of students gave answers such as 'shampoo' and 'soap', while only 1.5% of students gave such answers in version 2. The inclusion of the photograph showing household products in version 1 seemed to lead to the inappropriate meaning of the word 'products' being more likely to dominate students' mental models of the question.

Obviously students need to be able to cope with the chemical meaning of 'products' in science but it is important that the visual resources used or the context chosen are not such that they make such errors more likely.

Question 12 – Balance

In version 1 of question 12 the text states that end B moves down when weights are attached to the beam, but the beam is illustrated as horizontal. In version 2, the beam's realistic position is shown. The aim was to investigate whether students attempting version 1 might overlook the textual information since the diagram appears to supply the answer.

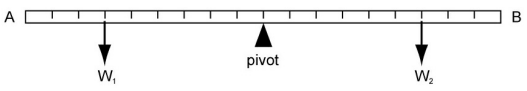
There was a statistically significant difference between marks on the two versions of the question (80.4% in version 1 and 98.8% in version 2 scored the mark, $F = 319.09$, $p < 0.01$). 16% of the students taking version 1 answered that the two weights were equal, suggesting that more attention had been paid to the diagram than the last sentence of the introduction.

Students clearly expected the diagram to reflect the answer. One student said, *That confused me because it's got the text saying one thing and the picture saying they're level*.

Question 12 : Version 1

12 A uniform beam AB is balanced at its midpoint on a pivot. Two weights W_1 and W_2 are then hung at equal distances from the midpoint of the beam.

When this is done, the end B moves down.



(a) Tick the correct statement.

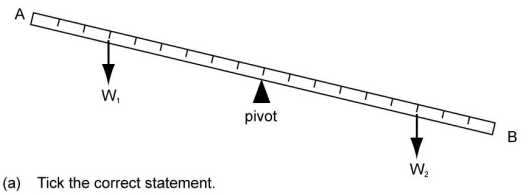
- ☐ W_1 weighs the same as W_2.
- ☐ W_1 is heavier than W_2.
- ☐ W_2 is heavier than W_1.

[1]

Question 12 : Version 2

12 A uniform beam AB is balanced at its midpoint on a pivot. Two weights W_1 and W_2 are then hung at equal distances from the midpoint of the beam.

When this is done, the end B moves down.



(a) Tick the correct statement.

- ☐ W_1 weighs the same as W_2.
- ☐ W_1 is heavier than W_2.
- ☐ W_2 is heavier than W_1.

[1]

Discussion

The analysis of the example questions in this study, along with others the authors have studied, suggest that two variables in particular play a decisive role in the effect of visual resources on the way examination questions are processed and answered. The first of these is the relative *salience* or prominence of the key elements. The most salient aspects of a question relative to the question as a whole are those that are emphasised in the student's construction of a mental model of the question. This degree of salience is to some extent determined by the idiosyncratic unconscious cognitive processing of students. However, as much as possible, the salience of certain aspects to students should match the aspects that the question setter considers important. Salience in examination questions can be increased by emphasising the feature in some way, for example by increasing its size, by using bold, or by including the element at the beginning of the question.

The student must also believe that the element is *relevant* to the answer. This decision is likely to be made at a more conscious level. One factor in determining this is past test experience, which provides expectations regarding under what circumstances visual resources are relevant. The type of illustration used and the context may also be of influence. For example, more technical diagrams, like the beam diagrams in question 12, were more likely to be used when answering than pictures or sketches.

The two variables were found to explain how the visual resources in the experimental test were used. For example, evidence from the interviews suggests that the size of the chicken nuggets' portion in question 5 was of average salience. However, relevance was deemed to be low, as students did not expect that portion size would be relevant in science.

Conclusion

The use of visual resources in examination papers can serve various positive purposes. However, the effects of illustrations are somewhat unpredictable and hence caution is required. It is important to ensure that, when used, visual resources are accurate and unambiguous. In addition, irrelevant information should not be included within a visual resource (except where selection skills are to be tested) as this may result in the wrong information being used. If a visual resource is not strictly needed in a question, the setter will need to balance the advantage that it may make the question seem less daunting against the possible risks that parts of the text may not be read thoroughly, or that a student may be led astray by an element in a visual resource that was not intended to be important.

Taking into account the salience and perceived relevance of visual resources can aid the prediction of their effects. The salience of aspects of visual resources will affect students' understanding of the question, and therefore the aspects that are key to the question should be the most salient ones in the question as a whole, whether they are presented in text or by illustration. If the key information is in the text, then any visual resource should support it rather than contradict or draw attention away from it.

Students have expectations relating to the type of visual resource used in a question, which may influence the kind of reasoning that they use. Evans and Over (1996) distinguish between two kinds of reasoning. What

can be called 'naturalistic reasoning' is innate, automatic and associative and is used in everyday functioning, whilst 'formal reasoning' is logical, controlled, reflective and learnt. It is the latter reasoning that examinations generally seek to assess. Evidence from questions 2 and 12 suggest that scientific diagrams are more likely to encourage formal reasoning, and naturalistic pictures are more likely to elicit naturalistic reasoning. Therefore the choice of type of visual resource also requires careful consideration.

This study constitutes a further stage in the collection of empirical evidence on the effects of features of examination questions on difficulty and validity. The information obtained from such research is used to inform training for question writers.

Further reading

The full report on this research can be found at http://www.cambridgeassessment.org.uk/research/confproceedingsetc

Acknowledgements

The question referred to as 'Question 2 – Statue' in this paper is reproduced by permission of QCA.

The question referred to as 'Question 5 – Children's meal' is adapted from GCSE Design and Technology: Food Technology 1460 paper 3, 1998 and reproduced with the kind permission of OCR.

The question referred to as 'Question 6 – Products' is adapted from GCSE Salters Science Double Award 1774 paper 1, 2000 and is reproduced with the kind permission of OCR.

The question referred to as 'Question 12 – Balance' is adapted from IGCSE Physics 0625 paper 2, 1999 and reproduced with the kind permission of CIE (University of Cambridge International Examinations).

We would also like to thank the teachers and students at the schools involved for their participation.

References

Biedermann, I. (1981). 'On the semantics of a glance at a scene.' In M. Kubovy and J. R. Pomerantz (Eds) *Perceptual organization*. Hillsdale, NJ: Erlbaum.

Evans, J. St. B. T. and Over, D. E. (1996). *Rationality and Reasoning*. Hove: Lawrence Erlbaum.

Fisher-Hoch, H., Hughes, S. and Bramley, T. (1997). 'What makes GCSE examination questions difficult? Outcomes of manipulating difficulty of GCSE questions': a paper presented at the British Educational Research Association Annual Conference. Cambridge: University of Cambridge Local Examinations Syndicate. Available from http://www.cambridgeassessment.org.uk/research/confproceedingsetc

Gernsbacher, M. A. (1990). *Language comprehension as structure building*. Hillsdale, NJ: Erlbaum.

Johnson-Laird, P. (1981). 'Mental models of meaning.' In A. K. Joshi, B. L. Webber and I. A. Sag (Eds), *Elements of Discourse Understanding*. Cambridge: Cambridge University Press.

Kennedy, J. M. (1974). *Psychology of picture perception*. San Francisco: Jossey-Bass.

Levie, H. W. and Lentz, R. (1982). 'Effects of text illustrations: A review of research', *Educational Communication and Technology Journal*, **30**, 195–232.

Ollerenshaw, A., Aidman, E. and Kidd, G. (1997). 'Is an illustration always worth ten thousand words? Effects of prior knowledge, learning style and multimedia illustrations on text comprehension', *International Journal of Instructional Media*, **24**, 3, 227–238.

Paivio, A. (1975). 'Imagery and long-term memory.' In A. Kennedy, and A. Wilkes (Eds), *Studies in Long Term Memory*. London: Wiley.

Peeck, J. (1987). 'The role of illustrations in processing and remembering illustrated text.' In D. M. Willows and H. A. Houghton (Eds), *The Psychology Of Illustration, Volume 1 Basic Research*. New York: Springer-Verlag.

Peeck, J. (1993). 'Increasing picture effects in learning from illustrated text', *Learning and Instruction*, **3**, 227–238.

Pollitt, A. and Ahmed, A. (1999). 'A new model of the question answering process': a paper presented at the International Association for Educational Assessment Annual Conference. Cambridge: University of Cambridge Local Examinations Syndicate. Available from http://www.cambridgeassessment/research/confproceedingsetc

Sarason, I. G. (1988). 'Anxiety, self-preoccupation and attention', *Anxiety Research*, **1**, 3–8.

Schnotz, W. (1993). 'Some remarks on the commentary on the relation of dual coding and mental models in graphic comprehension', *Learning and Instruction*, **3**, 247–249.

Stewart, J. H., Van Kirk, J. and Rowell, R. (1979). 'Concept maps: A tool for use in biology teaching', *The American Biology Teacher*, **41**, 171–175.

Weidenmann, B. (1989). 'When good pictures fail: an information-processing approach to the effect of illustrations.' In H. Mandl and J. R. Levin (Eds), *Knowledge acquisition from text and pictures, Advances in Psychology 58*. Amsterdam: Elsevier.

Winn, W. (1987). 'Charts, graphs, and diagrams in educational materials.' In D. M. Willows and H. A. Houghton (Eds), *The Psychology of Illustration, Volume 1 Basic Research*. New York: Springer-Verlag.

Winn, W. (1989). 'The design and use of instructional graphics.' In H. Mandl and J. R. Levin (Eds), *Knowledge acquisition from text and pictures, Advances in Psychology 58*. Amsterdam: Elsevier.

Gold standards and silver bullets: assessing high attainment

John Bell Principal Research Officer, Evaluation & Validation Unit

One of the challenges facing those involved in the assessment and selection of high attainers is the fact that so many students get the same high grades (in measurement theory this is referred to as a lack of discrimination). For students who are concerned about their future opportunities the assessment of high attainment and the lack of discrimination at the top end of the ability range can be crucial. For example, the proportion of successful applicants to Cambridge who gained three grade As at A-level (excluding General Studies) rose to 93% (Cambridge University, 2005). However, of the 8,026 applicants who met the 3As criterion, 5,325 were not accepted (note the preference at Cambridge is for depth of knowledge in a small number of relevant subjects rather than breadth of knowledge in more subjects, so taking more than three A-levels is not necessarily regarded as desirable). As Geoff Parks, Director of Admissions for the Cambridge Colleges, noted, "Cambridge would prefer applicants thinking of stretching themselves, having chosen a coherent set of A-levels, to do so by stretching themselves 'vertically' by taking one or two Advanced Extension Awards rather than 'horizontally' by taking a further A-level". (http://www.cam.ac.uk/admissions/undergraduate/info/statements/pallisreview.html)

Of course, not every A-level student with three grade As applies to Cambridge. In 2004 there were 21,101 eighteen-year-olds with at least three grade As at A-level (excluding General Studies). This number represents 3.5% of all eighteen-year-olds and 7.9% of the eighteen-year-olds with at least one A-level result. This is one of the problems. From the perspective of admissions to elite courses the number is high, but from the perspective of assessment provision as a whole it is small. For individual subjects the problem is even worse. The numbers and percentage entries for various A-levels are given in the Table below. When the A-level content is central to a higher education course, then there may be many more applicants with a grade A in the required subject than places.

A-level Subject	*Number with grade A*	*% with grade A*
Biology	11,511	24.8
Chemistry	11,289	31.8
Physics	8,217	29.6
Mathematics	20,093	40.0
Further Mathematics	3,433	60.1
Geography	8,346	24.7
History	10.723	25.0
Economics	5,476	31.7
English	12,846	25.9

Source: 2004 Inter-Awarding Body Statistics

A-levels were introduced in 1951 for the purpose of determining who should be admitted into a limited number of university places. The situation is different now. Government policy is to increase the numbers in higher education and one consequence is that a much larger number of students have been allocated to the same number of grades. This means that there are more applicants for higher education with identical levels of performance when classified by grade combinations.

Something that happens in every field of human endeavour, from being a mechanic to a rocket scientist, is that casual observers are all too willing to offer simple solutions to problems, forgetting H.L. Menken's metalaw, 'For every human problem, there is a neat, simple solution; and it is always wrong'. In the case of education, the simple solutions are often wrong because they often fail to solve the problem. To many people, the objective seems to be partitioning the highest performers so that there is a smaller, more manageable group from which to select. However, the problem is not one of partitioning but one of measurement. The criteria that need to be considered are:

- What is meant by high attainment and who has it?
- Does the assessment predict future performance?
- What are the implications for fair access?
- What impact will this have on learning in schools and colleges?

Research is needed to understand what exactly is meant by high attainment and the answer may differ for different subject areas. When deciding on a method for identifying high attainment it is necessary to consider what exactly is being rewarded. The usefulness of selection tests can be summed up by the title of Dorothy Field's song, 'It's not where you start, it's where you finish'. The objective of selection procedures is to identify those applicants most likely to succeed. Of course, it has to be recognised that a selection test can only measure some determinants of performance. Implications of the assessment for fair access need to be considered to ensure that the assessment does not create unnecessary barriers to admissions. Students have to have equality of opportunity in demonstrating their level of attainment and their potential to succeed in higher education. For high attaining candidates the stakes are high and this means that 'open' assessments, such as coursework and dissertations, might not be perceived to be fair. Finally, a high stakes test is bound to have an impact on what happens in schools and colleges so it is important to recognise the wide range of purposes of school examinations besides selection for higher education.

At high levels of attainment the issue is not usually one of standards, rather it is one of selecting the best students. It is not usually the case that the applicants will not be able to cope with the demands of the course but that some will achieve more. The selection process is not about meeting a standard but about being the most suitable candidate. At lower levels of attainment, standards are more of an issue and involve

the idea of minimal competence. It is also useful to recognise that particular A-levels may have different uses in different selection circumstances. These can be interpreted as evidence of attainment and evidence of potential. In the former, the A-level content is relevant and is used as a foundation for further learning and in the latter case, it is used as evidence that the candidate can cope with learning in a new area. In practice, this dichotomy is an oversimplification and both are relevant but the weight attached to each varies. The consequence of this is that responses to changes to the examination system will vary depending on who is using the examination results and for what purpose.

One result of the increased numbers with the same combination of grades is that there is either a lack of compensation (when a good performance in one subject is allowed to make up for a poor performance in another subject) or a reliance on performance in subjects of low relevance. For example, a student takes History, English, and Music A-levels and applies to do History at University. The offer is History A, English A, Music B but the student gets History A, English A, Music C. The student cannot compensate for the performance in music by showing an exceptionally high level of attainment in History. Who is more likely to be the better student, one who would have got an A++ History grade if one existed and a grade C in music, or a borderline History grade A and a grade B in Music? This issue of compensation is the major limitation of the new UCAS tariff (http://www.ucas.com/candq/tariff/index.html). There is no ceiling on the number of points than can be acquired and so compensation does not operate in a useful way. This means that a candidate who obtained only three grade As at A-level and had no other qualifications would have the same points as a candidate with three grade Cs at A-level and three grade Cs at AS. The latter candidate has not demonstrated any high level skills associated with a grade A. The two candidates are not equivalent.

The use of tariffs is further limited because it fails to take into account the requirements of particular HE courses. Hence, tariffs should not be used for evaluating whether the admissions process is fair (in terms of equal opportunities). For example, when considering those candidates with three grade As at A-level, including A-levels required for an individual course (i.e. most of the pool of suitably qualified candidates for a Cambridge course), the percentage of candidates meeting the criteria and attending independent (public) schools varies from 36.8% for courses requiring mathematics and physics to 51.4% for those requiring a language A-level.

Lotteries

Using the criteria listed above it is possible to consider various options that have been proposed to select high attaining candidates. One solution that has been proposed is the use of lotteries. Nothing is being measured and they are based on the assumption that all students have an equal probability of success. This is simply not the case. Within grade A there is a considerable variation in performance on the assessments described below and this performance is related to success. In terms of fair access, differences between a statistical concept of fairness (having an equal probability of being selected) and the real word 'fairness' (may the best person win) is an issue. In America where lotteries have been used, objections have been raised when the weaker students in the same institution (so resourcing and background are not issues) are selected in preference to the stronger. Lotteries are not an acceptable solution.

Marks

Another popular common suggestion is to use marks. The first difficulty is that the raw marks that are awarded to candidates are not really useful. A-level examinations are made up of modules and the combinations of modules vary from candidate to candidate. This can be the result of option choices or of taking the modules in different examination sessions. With a few exceptions, it is very difficult to construct examination modules of equal difficulty from session to session without pre-testing. This means that the raw marks have to be converted to uniform marks so that a grade A is always worth 80% of the marks on any module and in any session. This Uniform Mark Scheme was designed to be an intermediate step in awarding grades. Unfortunately, the system is designed to be fair to candidates close to the grade A boundaries and has potential problems for higher levels of performance. There is also a problem with exactly what is being measured. To ensure that the rate of exchange between raw and uniform marks is the same just above and below the highest grade, a cap is introduced. In some circumstances, the cap is lower than the maximum raw marks and maximum UMS marks are awarded to all candidates with raw marks equal to or above the cap (all but extremely erratic candidates obtain a grade A even if one or more of their modules is capped). Unfortunately, this capping process is an issue at higher levels of attainment.

Additional grades above A-level

Obviously the above property of UMS marks also affects the introduction of additional grades above A-level. There is also the issue of what exactly is being measured. There are two approaches that can be adopted. The first involves setting new boundaries on existing papers and the second involves adding additional material. However, there are difficulties associated with both of these alternatives. In the first case, the problem is that the examinations are not necessarily designed to test higher level skills above grade A. In investigating the feasibility of introducing additional grades at A-level, it was found that in some subjects it was considered possible to do this with existing examinations but in others it was felt that additional material would be essential. In these circumstances, it was felt that it would reward conscientious and careful students but would not identify those with higher order skills. However, adding harder material to the examinations is not straightforward because it alters the measurement characteristics of the examination as a whole. This is manifested either in very low grade E boundaries, a compression of the mark range between the A and E boundaries, or a lengthening of the examination, increasing the assessment burden. It should also be noted that only a small percentage of the candidates would be likely to complete the task satisfactorily because of its increased difficulty.

It is possible to investigate some consequences of extra grades using existing A-levels. For example, introducing A+ and A++ grades at equally spaced intervals gives the following results. For one OCR mathematics specification, 74% of candidates attended state schools but only 63.1% of those obtaining a grade A attended state schools. For the hypothetical A+/A++ grade the percentages are 55.3% and 49.1% respectively. Similar patterns were found for a range of other A-level subjects with pupils from independent schools increasingly represented in the higher range of the mark distribution.

Module grades

Another option is to use the grades obtained on each module. To be fair this would require all A-levels to have the same module structure. Even if this were the case then the tendency is to reward consistency. For example, consider two candidates with the following module grades (in lower case) and UMS marks (in brackets):

Candidate X: a(80), a(80), a(80), a(80), a(80), a(80)

Candidate Y: a(100), a(100), a(100), a(100), a(100), b(79).

If only the module grades are known and used then candidate X seems better but the UMS marks indicate that Y is almost certainly the better candidate. Obviously this is an extreme example but deciding whether a consistent performance is better than an erratic performance is debatable as it depends on the circumstances. Consistency is important for airline pilots as being brilliant at take-offs but useless at landings does not make a satisfactory pilot, but the history of the arts abounds with individuals classed as great for some of their work that was brilliant, even though much of what they did was less outstanding.

Additional assessments

There are other alternatives involving additional assessments. These can be grouped into three types:

- Subject specific examinations, e.g. Special papers and Advanced Extension Awards;
- HE course specific, e.g. the BioMedical Admissions Test (BMAT);
- General tests of high order skills, e.g. thinking skills assessments.

These tests have their own particular advantages and disadvantages. All of them have the advantage that they can be targeted solely for high attainers but there are also disadvantages in that they lead to additional costs and an increased assessment burden on candidates.

Subject specific examinations have the advantage that they are based on a particular subject area so that the measured issues can be addressed without compromise. However, there are access issues. In 2004 the uptake of Advanced Extension Awards was relatively low, from just 2 candidates (Irish) to 1,501 candidates (English). Universities cannot use them for making admissions' offers unless they are available in all schools so that they can be made a requirement. It can also be argued that such examinations favour schools and colleges that are either large, highly selective or well resourced because such institutions can provide the most effective support for candidates entering such examinations.

Another option takes the form of tests designed for admission to specific courses. This has the advantage that only relevant skills and knowledge are assessed. This can be a subset of the content of an A-level specification and can also include skills not directly assessed at A-level. In addition, all applicants can take them whether they enter A-level examinations or not and so they provide a common reference point for making decisions. These tests provide a way of assessing the potential of students whose ability might not be reflected in their grades. The objections relate to the extent that performances can be improved by coaching. This is a difficult issue. For example, it is accepted that the skills used in the BMAT will improve with familiarity and practice. However, it can be argued that these skills are really worthwhile, useful in many walks of life, and very important for success in higher education, and in this case it is possible for anyone to practice the skills involved with the help of freely and publicly available materials which are listed on the BMAT website.

Finally, there are tests that measure skills that are not directly assessed or not assessed by general qualifications. These skills may be important to success in higher education but it is important that the predictive validity is established. There has been some UK research into this issue (MacDonald et al, 2001a, 2001b). In a trial of the American College Board's SAT it was found that a different subset of high attaining candidates would be identified by the SAT compared with A-level. However, the study could not address whether these candidates would perform better than those identified by A-levels. The issue of predictive validity was not addressed.

One problem with experimental research is that it is low stakes. The outcome of the test has no impact on the future of the candidates taking it. The schools of the students in the experiment were not making any effort to prepare candidates for the test. If an aptitude test such as the American SAT were used for admissions, then this situation would change. For example, the BBC correspondent, Mike Baker, reported that "At Lafayette High School in Williamsburg, Virginia, I sat in on an 'SAT preparation' class. It's a sign of how important the SAT is in a teenager's life that this runs for 90 minutes a day for a whole semester. Piled high in the corner were some of the many preparation text books available on the market." (http://news.bbc.co.uk/1/hi/education/3304459.stm). It is for this reason that it is important that any general admissions test developed for higher education admissions should be developed to assess aptitudes that are educationally important and have long term benefits. Extensive research shows that coaching on aptitude tests has a small effect (Powers and Rock, 1999). However, this research was carried out to counter claims of coaching made by commercial providers. It does not mean that educational experiences do not influence the SAT score. For example, one of the College Board's researchers, Howard T. Everson, with a colleague, Roger E. Millsap, from Arizona State University (2004), investigated influences on SAT performance. They found that family background, learning opportunities in and outside the school curriculum and school characteristics influenced the SAT score.

Conclusions

In this paper, a number of options for assessing a small but important subset of candidates have been considered. UCLES has wide experience of all of these options and UCLES' researchers have conducted and will conduct many research projects that investigate the effectiveness of them. In particular, new methods are being developed to investigate the crucial factor of predictive validity. This research will be described in a future issue.

Whatever method of assessment is used, its effectiveness depends on how well it predicts future behaviour. This varies with circumstances so there is no simple gold standard that always identifies the best candidates in all circumstances. Neither is there a silver bullet – a test that measures the candidate's potential uninfluenced by an individual candidate's education experiences and personal circumstances. It is unreasonable either to expect this or to claim it of any educational assessment. However, it is important that assessments designed for this purpose do not add any biases and that they identify the candidates

most likely to succeed. In addition, any preparation for the test should have a beneficial effect on the candidate, equipping them with skills that they will need as they progress through life.

References

Boyle, C. (1998). 'Organisations selecting people: how the process could be made fairer by the appropriate use of lotteries (with discussion)', *Journal of the Royal Statistical Society: Series D: The Statistician*, **47**, 2, 291–322.

Everson, H.T., and Millsap, R.E. (2004). 'Beyond individual differences: exploring school effects on SAT scores', *Educational Psychologist*, **39**, 3, 157–172. With correction **39**, 4, 261–261.

McDonald, A.S., Newton, P.E., Whetton, C. and Benefield, P. (2001). *Aptitude Testing for University Entrance: A literature review*. Slough: NFER.

McDonald, A.S., Newton, P.E. and Whetton, C. (2001). *A pilot of aptitude testing for University Entrance*. Slough: NFER.

Powers, D.E., and Rock, D.A. (1999). 'Effects of coaching on SAT I: Reasoning Test Scores'. *Journal of Educational Measurement*, **36**, 2, 93–118.

AUTOMATIC MARKING

Automatic marking of short, free text responses

Jana Z. Sukkarieh[1], **Stephen G. Pulman**[1] and **Nicholas Raikes**[2]

Introduction

Many of UCLES' academic examinations make extensive use of questions that require candidates to write one or two sentences. With increasing penetration of computers into schools and homes, a system that could partially or wholly automate valid marking of short, free text answers typed into a computer would be valuable, but would seem to pre-suppose a currently unattainable level of performance in automated natural language understanding. However, recent developments in the use of so-called 'shallow processing' techniques in computational linguistics have opened up the possibility of being able to automate the marking of free text without having to create systems that fully understand the answers. With this in mind, UCLES funded a three year study at Oxford University. Work began in summer 2002, and in this paper we introduce the project and the information extraction techniques used. A further paper in a forthcoming issue of *Research Matters* will contain the results of our evaluation of the automatic marks produced by the final system.

Uses for automatic marking

UCLES' traditional strength is in high stakes assessments that lead to qualifications. As more of our customers move to computer based assessments, an initial application of automatic free text marking in a high stakes context is as a quality control check on human marking, increasing the speed and efficiency of our quality control process. Every short, free text answer[3] could be marked both by computer and human markers, with any differences being resolved by a second human marker. Over time, as the capabilities and limitations of automatic marking became better understood, the proportion of answers marked by both computer and human could be reduced, with human marking targeted on the hardest to mark questions and on reviewing automatic marks that appear anomalous.

In the short term, however, the real opportunity for automatic free text marking is in low stakes tests. Many teachers and students use questions from our past papers, and we would like to be able to offer them an automatic marking service covering the free text questions as well as the 'objective' ones.

The challenge

Raikes and Harding (2003, p.270) state that an item's suitability for automatic marking depends on how near it can be placed to the objective end of what they call the objective-subjective continuum. The continuum is defined by the 'resolution' – the specificity and comprehensiveness – of an explicit marking guide that specifies how answers should be processed and marked. Traditionally, high resolution guides have been generated by greatly constraining the answers that students may give, as in multiple choice tests. More recently, attention has focussed on techniques for generating what are in effect high resolution marking guides for more open-ended item types, shifting them towards the objective end of the continuum where they may be automatically marked without affecting their validity.

In our automatic marking project we were concerned with marking short, factual answers varying in length from a few words up to around five lines, taken from GCSE biology examinations, where answers were marked for their correct content. The challenge was in coping with the myriad and sometimes unconventional ways in which credit-worthy answers were expressed, and the many mistakes in grammar and spelling found in some answers that nevertheless contained more or less the right content. Standard syntactic and semantic analysis methods would have been difficult to use, and even if we had fully accurate syntactic and semantic processing, many answers contained features that require a degree of inference that is beyond the state of the art. For example, in a question concerning asexual reproduction, a human marker inferred that

1. Computational Linguistics Group, Centre for Linguistics and Philology, Walton Street, Oxford OX1 2HG, United Kingdom. Email: first.lastname@clg.ox.ac.uk

2. Senior Research Officer, Evaluation & Validation Unit

3. Of course, completely objective items – multiple choice items and the like where every possible answer can be predicted – will be marked solely by computer.

a student who wrote *you do not have to wait until spring* meant to say *asexual reproduction can be done at any time*, a statement worth a mark according to the mark scheme. We have also found that students sometimes use a negation of a negation for a positive, as in *won't be done only at a specific time*, written for the same question. Contradictory or inconsistent information must also be detected, such as the inconsistent scientific information contained in the student statement *identical twins have the same chromosomes but different DNA*. These circumstances conspire to make the task too challenging for deep processing at present and so we decided to trade accuracy for robustness and investigate shallower 'information extraction' techniques, since they do not require complete and accurate parsing and are relatively robust in the face of ungrammatical and incomplete sentences.

Information extraction in a nutshell

Information extraction (IE) techniques pull out pertinent information from a partially syntactically analysed text by extracting those bits that match a set of domain-specific patterns typically built from training data. In our case, the training data are a sample of human marked answers – some human marking is necessary for setting up automatic marking – and the mark scheme, and a pattern is essentially all the paraphrases discovered for a particular entry in the mark scheme. The patterns include linguistic features as well as keywords.

Patterns

We wrote our initial patterns by hand, but have worked on a tool to take most of the tedious work out of this task. We base the patterns on recurring head words or phrases found in the training data, with syntactic annotation where necessary.

Consider the following six example training answers, which were written in response to the part question

Explain what has caused these two twins to be identical:

- the egg after fertilisation splits in two
- the fertilised egg has divided into two
- the egg was fertilised it split in two
- one fertilised egg splits into two
- one egg fertilised which split into two
- one sperm has fertilised an egg that split into two

These are all acceptable paraphrases of an answer given in the mark scheme as *They are formed from the same fertilised egg/same embryo*, and they and similar variants are captured by a pattern like:

singular_det + <fertilised egg> +{<split>; <divide>; <break>} + {in, into} + <two_halves>, where

<fertilised egg>	=	NP with the content of 'fertilised egg'
singular_det	=	{the, one, 1, a, an}
<split>	=	{split, splits, splitting, has split, etc.}
<divide>	=	{divides, which divide, has gone, being broken...}
<two_halves>	=	{two, 2, half, halves}

It is sometimes essential that the patterns incorporate the linguistic knowledge our syntactic analyser can generate at the moment, namely part-of-speech tags, noun phrases and verb groups. In the above example, the requirement that <fertilised egg> is a noun phrase (NP) will exclude something like *one sperm split in two and fertilised more than one egg* but accept something like *an egg which is fertilised*....

System architecture

The Student's View

Figure 1 is an annotated screenshot from a trial test that uses our automatic marking engine. Students enter their answers in the box at the bottom and may optionally click the 'check spelling and typing' button, which identifies any unrecognised words and suggests alternatives. We currently permit students to edit their answers as much as they wish without penalty. When they are happy with their answers to all the questions, students click the 'Finish test' button which submits their answers for marking. Students receive their item level marks together with an indication of what the marks were awarded for. The system can easily be reconfigured to, for example, provide marks and feedback after each question attempt, rather than at the end, depending on the context in which it is being used.

Figure 2 is a schematic diagram of how the marking system works. In this case the answer *When the caterpillars are feeding on the tomato plants, a chemical is released from the plants* is fed into the syntax analyser, which tags the different parts of speech (POS) and identifies chunks of text that represent noun phrases (NPs) and verb groups (VGs). The analyser makes use of a general lexicon, derived from the *Wall Street Journal* and the British National Corpus, and a specialised lexicon, derived in this case from a GCSE biology textbook and other specialised vocabulary encountered in the training data. The tagged and chunked answer then goes into the pattern-matching and marking system, where it is matched, if possible, with the patterns constructed from the training data. Marks and justifications are issued according to the scores and justifications pre-determined for the patterns matched and the rubric of the mark scheme.

Preliminary indications of marking accuracy

We will give the results of a larger and more detailed trial in a future article in *Research Matters*, but results from a preliminary trial involving nine part-questions are given in Table 1. The marking patterns were manually written using a training set of around 200 answers marked once by a human examiner, and the table gives the results when these patterns were tested on a further 60 answers that had not been seen by the developers until after the patterns were written. Note that the full mark for each question ranges between 1 and 4 – that is, the number of correct points required ranged from 1 to 4; there is a one-to-one correspondence between a correct point and a mark.

Column 3 records the percentage agreement between our system and the marks assigned by a human examiner. Sometimes humans make mistakes, however, and column 4 reflects the degree of agreement between the marks awarded by our system and those that would have been awarded by following the marking scheme consistently. Notice that agreement is correlated with the mark scale: the system appears less accurate on multi-point answers. We adopted an extremely strict measure of agreement, requiring an exact match.

Figure 1: Annotated screenshot of a trial automatically marked test

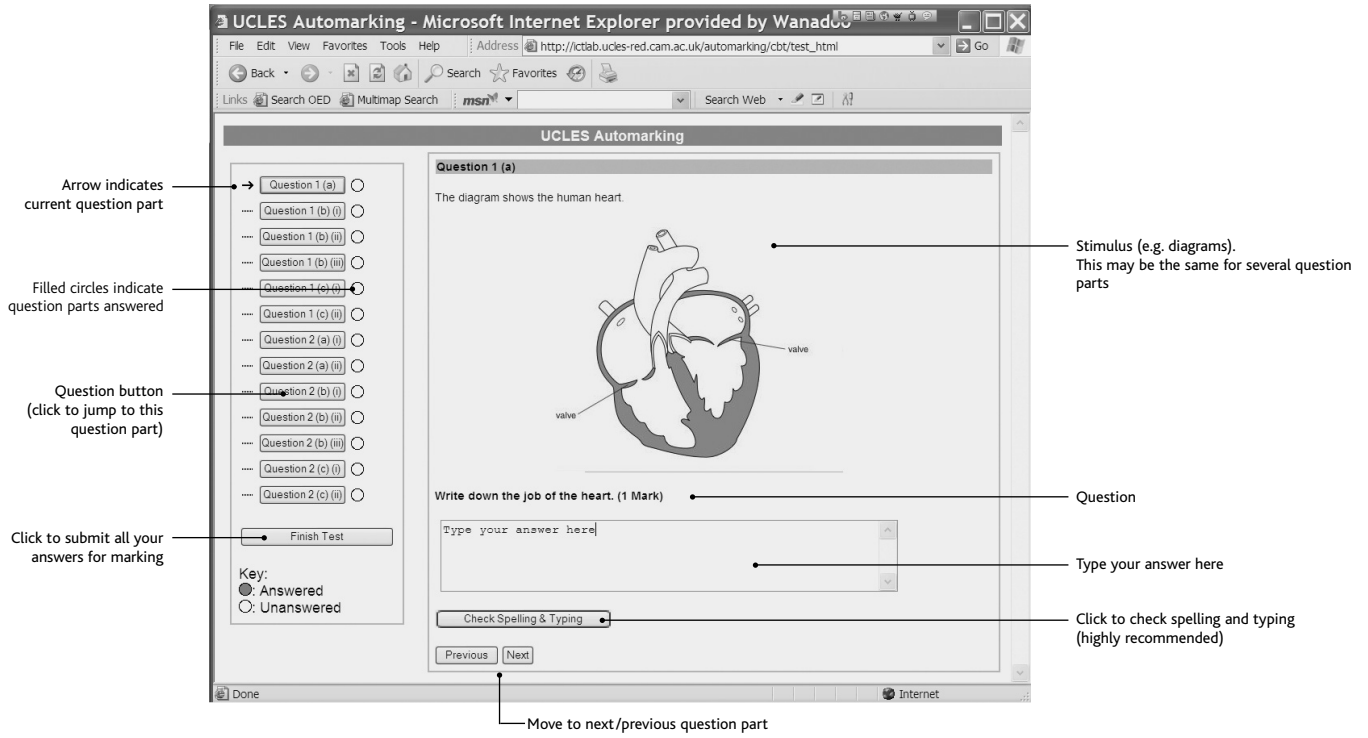


Figure 2: Schematic diagram of the marking system

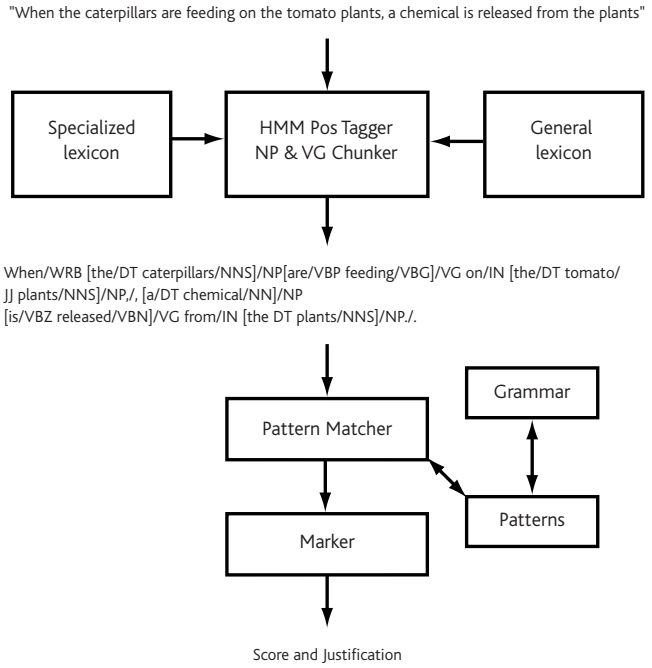


Table 1: Results for a preliminary evaluation of automatic marks compared with human marks, using manually written patterns and an Information Extraction approach

Question	*Full Mark*	*% Examiner Agreement*	*% Mark Scheme Agreement*
1	2	89.4	93.8
2	2	91.8	96.5
3	2	84.0	94.2
4	1	91.3	94.2
5	2	76.4	93.4
6	3	75.0	87.8
7	1	95.6	97.5
8	4	75.3	86.1
9	2	86.6	92.0
Average	—	85.0	92.8

Others' work

Several other groups are working in this area. The most prominent systems are C-Rater, developed by Leacock et al. (2003) at the Educational Testing Service (ETS), the IE-based system of Mitchell et al. (2003) at Intelligent Assessment Technologies, and that at Carnegie Mellon University described by Rosé et al. (2003). The four systems (these three and ours) are being developed independently, yet it seems they share similar characteristics. Commercial and resource pressures currently make it impossible to try these different systems on the same data, and so performance comparisons are meaningless.

Practical limitations and extensions

It takes around a day and half for a developer to discover and write the patterns manually for a new question, and we require around 200 definitively marked answers for pattern writing. We are currently evaluating whether a non-specialist programmer without experience in computational linguistics can do this task as effectively. We have also done some quite promising work on semi-automating the pattern writing process to make it quicker and less labour intensive – see Sukkarieh et al (2004) and Sukkarieh and Pulman (2005) for more information and results. Alternative, non-IE machine learning approaches have also been trialled with varying degrees of success –Sukkarieh and Pulman (2005) give details and results.

Conclusion

We have introduced our automatic marking project and described the information extraction techniques used and how we have applied them. Initial results are encouraging, with automatic marks correct 93% of the time on average. We will present the results of a more wide-ranging evaluation in a future edition of *Research Matters*.

References and further reading

Leacock, C. and Chodorow, M. (2003). 'C-rater: Automated Scoring of Short-Answer Questions'. *Computers and Humanities,* **37**, 4.

Mitchell, T., Russell, T., Broomhead, P. and Aldridge, N. (2003). 'Computerized marking of short-answer free-text responses': paper presented at the 29th annual conference of the International Association for Educational Assessment (IAEA), Manchester, UK.

Raikes, N. & Harding, R. (2003). 'The horseless carriage stage: replacing conventional measures', *Assessment in Education*, **10**, 3, 267–77.

Rosé, C. P., Roque, A., Bhembe, D. and VanLehn, K. (2003). 'A hybrid text classification approach for analysis of student essays'. In *Building Educational Applications Using Natural Language Processing*, 68–75.

Sukkarieh, J. Z. and Pulman, S. G. (2005). 'Automatic Short Free-Text Answer Marking', *Natural Language Engineering*, (under review).

Sukkarieh, J. Z., Pulman, S. G. and Raikes N. (2003). 'Auto-marking: using computational linguistics to score short, free text responses', paper presented at the 29th annual conference of the International Association for Educational Assessment (IAEA), Manchester, UK.

Sukkarieh, J. Z., Pulman, S. G. and Raikes N. (2004). 'Auto-marking 2: An update on the UCLES-Oxford University research into using computational linguistics to score short, free text responses': paper presented at the 30th annual conference of the International Association for Educational Assessment (IAEA), Philadelphia, USA.

Research News

The Cambridge Assessment Conference:

A Question of Confidence – Maintaining Trust in National Assessment Systems

This conference takes place on 17 October 2005 at Robinson College, Cambridge and will be run by Cambridge Assessment Network. The Network is part of Cambridge Assessment, formerly known as the UCLES Group, a not-for-profit body recognised throughout the world for its professional, rigorous and high-standard assessments.

The conference will address a current issue about public examining and assessment. Leaders from many areas within education are being invited, including senior people from schools, colleges and universities and those working for national and local education bodies, professional organisations, political parties, the media, awarding bodies and employers' organisations.

The four main speakers and the titles of their papers are:

- Baroness Onora O'Neill, Principal of Newnham College, Cambridge
  Assessment and Public Accountability
- Professor Barry McGaw, Director for Education, OECD
  Assessment in Education
- Dr Nicholas Tate, Director General of the International School of Geneva
  The School, the Parents and the Exam Boards
- Professor Alison Wolf, Sir Roy Griffiths Professor of Public Sector Management, King's College, London
  What can and should we be measuring?

There will also be five discussion seminars on the following topics:

1. Maintaining standards
2. Teacher assessment
3. Vocational qualifications
4. Information technology
5. Working with the mass media.

For further information and an application form please email thenetwork@cambridgeassessment.org.uk.
Tel: 01223 553846;
www.cambridgeassessment/events

Other conferences and seminars

Colleagues from within the Assessment Directorate have attended several conferences and seminars in recent months.

Joint British Educational Research Association/Local Government Association Research Conference, April 2005

Martin Johnson of the Research Programmes Unit gave a keynote presentation entitled *On-line assessment: the impact of mode* at Local Government House, London. The conference was attended by representatives from LEAs and universities with an interest in assessment research. Other presentations were given by Professor Richard Daugherty who spoke on *Developing assessment policies: a role for academic research?*, Professor Wynne Harlen *Use of teachers' judgements in summative assessment*, and Tonnie Richmond *The uses and abuses of value added data*.

General Teaching Council (GTC) seminar: What kind of assessment model best enhances teaching and learning? May, 2005

Presentations were given by Professor Wynne Harlen, Professor Richard Daugherty and representatives from the Assessment Reform Group, QCA, NAA and Becta. The discussions were intended to further the development of the GTC's assessment policy and the advice it gives to the Government.

The British Psychological Society's Quinquennial Conference, April, 2005

The keynote speaker on one of the days was Robert Sternberg of Yale University who spoke on *Redefining Elementary through Postgraduate Education through the Theory of Successful Intelligence*. This was about his theories on learning styles. Of particular interest was the explanation of research into how an admissions test of items testing each of the three types of intelligence increased quite substantially the amount of variance in College/University grades that could be explained by the American SAT test and High School grades alone.

GTC conference: The future role of e-assessment

The session included presentations and discussions aimed at gathering the opinions of various educational agencies in order to inform the GTC's future policy development and feedback to government. Martin Ripley (Head of the e-assessment team, QCA) reported on the QCA vision for e-assessment over the next 5 to 10 years and on the latest news of the KS3 ICT online test pilot. A representative from the DfES' e-learning strategy unit outlined the department's goal for a unified holistic system for all involved in education.

In June the Innovation in Assessment and Learning Unit was invited to contribute to the *Innovation in Teaching and Learning Workshop* organised by the Learning and Skills Development Agency. The focus was on the vocational sector. As well as members of the Steering Group of the LSDA's Innovation Project, the event was attended by senior members of the DfES Standards Unit, the QCA, the Institute of Education and FE Colleges. Several themes emerged as being of great interest to delegates including e-portfolios, e-communities, collaborative learning and peer assessment.

Cambridge Assessment
1 Hills Road
Cambridge
CB1 2EU

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