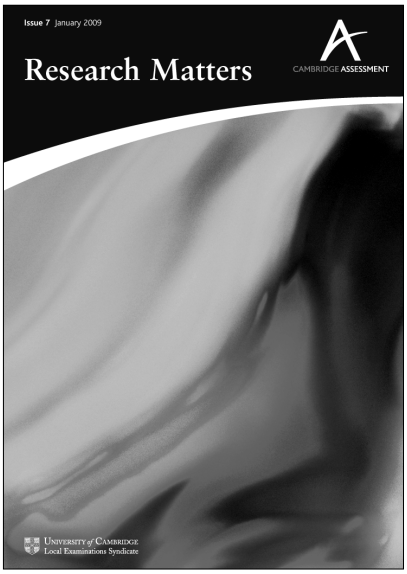
Issue 7 January 2009

Research Matters

CAMBRIDGE ASSESSMENT

University of Cambridge
Local Examinations Syndicate



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

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

Research Matters : 7

A CAMBRIDGE ASSESSMENT PUBLICATION

Foreword

While the contributions in this issue each provide insights into diverse matters in assessment one article in particular warrants special attention. Gray's and Shaw's contribution on 'de-mystification' of the Uniform Mark Scheme (UMS) is a valuable example of researchers 'drawing breath and taking stock' of processes and concepts located deeply within procedures for administering contemporary public examinations. Both authors benefit from being immersed in the day-to-day processes of administering examinations as well as contributing to the sometimes fraught and always pressured processes of developing new ones. In Robert Wood's 1991 book for UCLES, *Assessment and testing – a survey of research*, he drew attention to the serious implications of the significant gap between public understanding of examinations and the actual characteristics and technical processes of assessment. Interestingly, increasing public understanding of assessment was a feature of the School Council's work in the 1970s, but despite intense pressure from a few lone voices, this was not supported as a key mission of its successor bodies – even to the present day. Surely, as assessment professionals, we are more accountable to candidates and the users of assessment data than this taciturn position suggests. The UMS is central to Advanced level qualifications in England in particular, yet it is widely misunderstood and the artefacts of its use even less understood. The article included here is thus crucial not only in opening up technical discussion, but is a vital contribution to increasing transparency and openness.

Tim Oates *Group Director, Assessment Research and Development*

Editorial

Most of the articles in this issue report on research that was presented at the British Educational Research Association (BERA) and/or the International Association for Educational Assessment (IAEA) conferences during the autumn of 2008. In the first article Sylvia Green summarises the two keynote presentations from the IAEA 2008 conference. In different ways they look to the future of assessment. Professor Robert Mislevy focussed on the implications of expertise research in the context of computer based testing and task design while Professor Dylan Wiliam explored the meanings of educational assessment and the challenges faced in the design of future systems.

The second article from Raikes, Scorey and Shiell explores a method to enable a greater range and number of educational professionals to contribute to decisions on grade boundaries. The research reported by Greatorex and Nádas considered whether using 'think aloud' methods to investigate assessment judgements compromises the authenticity of the thought processes involved. This is a methodological concern that has been expressed and debated widely and their work adds to this field of knowledge with encouraging results. The aim of the research on emotional intelligence by Vidal Rodeiro, Bell and Emery was to investigate whether relationships exist between the affective domain and progress in school. The Evaluation and Psychometrics team marked Cambridge Assessment's 150th anniversary by looking back at question papers over the years. The project is summarised in 'Assessment instruments over time'. Elliott and Johnson's article provides a detailed analysis of the nature of the spelling errors identified in the 'Aspects of writing' project. The aim was to establish whether certain spelling errors were particularly common and how they related to spelling conventions, as taught in schools.

In the first of the articles from OCR and Cambridge International Examinations (CIE) Gray and Shaw attempt to demystify the UMS employed in the examination system. They discuss some of the practical challenges posed by the calculation of grades for unitised specifications. The second article, from Shaw, outlines the CIE research agenda from routine operational procedures to more full-scale experimental investigations.

Statistical reports are listed for information and are based on the annual national – level examination databases for pupils in England.

Sylvia Green *Director of Research*

Keynote presentations to the International Association for Educational Assessment (IAEA) 2008 Annual Conference

Sylvia Green Research Division

The 34th IAEA annual conference, hosted by Cambridge Assessment, took place in Robinson College, University of Cambridge from September 7th to 12th. The main conference theme was *Re-interpreting Assessment: Society, Measurement and Meaning*. The conference was the largest IAEA conference ever with around 500 delegates from 58 countries; 130 papers and 8 posters were presented. The highlights of the event were the two keynote presentations by Professor Robert Mislevy and Professor Dylan Wiliam.

PROFESSOR ROBERT MISLEVY:

Some implications of expertise research for educational assessment

The first keynote was presented by Professor Robert Mislevy, Professor of Measurement and Statistics at the University of Maryland. He was previously Distinguished Research Scientist at ETS and is a member of the National Academy of Education. He has been president of the Psychometric Society, and received career awards from the National Council on Measurement in Education and the American Educational Research Association. His research applies developments in technology and cognitive psychology to practical problems in assessment. In his address Mislevy focussed on the implications of expertise research for educational assessment commenting that developments in psychology and technology had led to exciting times in assessment. He provided insights from his research and their implications for assessment design. His descriptions of complex cognitive processes were presented in a way that was accessible to his audience and he provided meaningful illustrations of the concepts he introduced. He outlined difficulties in the contexts of cognitive processing limitations and knowledge, describing expertise as 'the circumvention of human processing limitations' (Salthouse, 1991). He also explained Walter Kintsch's theory of reading comprehension where relevant patterns from long-term memory may be activated in some contexts and not in others. In this theory the writer depends on many patterns and conventions that have developed over hundreds of years through the interactions of billions of people in the form of the letters, the syntax, and the words of the language itself. Mislevy explained that the physical and social context, what we have been working on, our purpose for reading, even the time of day, can influence our comprehension. He went on to draw on Kintsch's theory in relation to expertise research.

In expertise research into cognitive task analysis experts and novices are compared in replicable conditions and questions are asked, such as, *What knowledge is needed? How is it represented? How is it used? What makes tasks hard?* He suggested that experts organise their knowledge effectively and that they perceive, understand and act in terms of fundamental principles rather than surface features (Chi, Feltovich and Glaser, 1981). He emphasised the importance of interaction with a situation and of external knowledge representation for information processing and cognition. A key question posed was – *How do you use improved understanding of the nature and acquisition of expertise to design and conduct assessments?* Other important questions were raised, for example, *What complex of knowledge, skills and other attributes should be assessed? What behaviours or performances should reveal those constructs? What tasks or situations should elicit those behaviours?* (Messick, 1994). He drew on a socio-cognitive perspective, looking at four important aspects of expertise:

- Organisation of knowledge
- Knowledge representations
- The importance of interaction
- Social aspects of expertise.

He went on to describe examples of computer assisted assessments with a range of design tasks and simulations and considered implications for task design and design patterns. The differences between experts and novices were set out in three contexts: using disparate sources of information; formulating problems and hypotheses; vocabulary and language usage. He concluded that insights from expertise research can improve the practice of assessment and support deeper learning and that doing so requires a deeper understanding of assessment design. He also suggested that suitable conceptual frameworks, tools and exemplars are now beginning to appear.

He concluded that assessment is – structuring situations that elicit evidence about students' thinking and acting in terms of patterns 'in our head' and that many of these patterns are social in their construction, acquisition and use. Some of the insights derived from cognitive psychology and developments in technology can be applied directly in familiar forms of testing while others need to be developed outside traditional and familiar practices, in technological environments. Mislevy's final comment was a challenge to assessment professionals: 'We, as the community with interests in learning and assessment, are moving to our next level of expertise.' This was a thought-provoking presentation that was well received by delegates. It set out some interesting thoughts that were referenced in different contexts throughout the conference.

PROFESSOR DYLAN WILIAM:

What do you know when you know the test results? The meanings of educational assessments

The second keynote address was presented on the final morning of the conference by Professor Dylan Wiliam, Deputy Director of the Institute of Education, London. In a varied career, he has taught in urban public schools, directed a large-scale testing programme, served a number of roles in university administration and pursued a research programme focussed on supporting teachers to develop their use of assessment in support of learning. He posed an intriguing question – 'What do you know when you know the test results?' In his presentation he explored the meanings of educational assessments and focussed on the conference theme, *Re-interpreting Assessment: Society, Measurement and Meaning*. He addressed a number of fundamental issues:

- The importance of (un)reliability
- Evolving conceptions of validity
- (Mis)uses of assessments for educational accountability
- Some prospects for the future.

He discussed the difficulties of using classical measures of reliability and of ensuring the accuracy of scores as well as the precision of grades. He proposed that a test is valid to the extent that it assesses what it purports to assess and that the key properties in terms of content validity are relevance and the extent to which the test is representative. He also discussed issues of content validity, criterion-related validity (concurrent and predictive) and construct validity and he identified three important issues related to validity:

- Validity is a property of inferences, not of assessments.
- The phrase 'a valid test' is therefore a category error ('like a happy rock').
- Reliability is a pre-requisite for validity.

He proposed that validity subsumes all aspects of assessment quality including reliability, content coverage, relevance and predictiveness, but not impact. He identified threats to validity in terms of inadequate reliability as – construct irrelevant variance and construct under-representation.

Wiliam then moved on to some practical applications. The first of these was in the area of school effectiveness. He asked, 'Do differences in student achievement outcomes support inferences about school quality?' In discussing this he referred to the threats to validity that he had previously identified. This led to a further exploration of the threat of construct irrelevant variance and construct under-representation. In outlining the social consequences of inadequate assessments Wiliam referred to the Macnamara Fallacy:

> *The first step is to measure whatever can be easily measured.*
> *(This is OK as far as it goes).*
>
> *The second step is to disregard that which can't be easily measured or to give it an arbitrary quantitative value.*
> *(This is artificial and misleading).*
>
> *The third step is to presume that what can't be easily measured really isn't important.*
> *(This is blindness).*
>
> *The fourth step is to say what can't be easily measured really doesn't exist.*
> *(This is suicide)*
>
> (Handy, 1994 p.219)

He also quoted Goodhart's Law – *All performance indicators lose their meaning when adopted as policy targets* – and he gave examples of this including inflation and money supply as well as national and provincial school achievement targets. He warned against the effects of narrow assessment because of the tendency to create incentives to teach to the test and to focus on:

- Some subjects at the expense of others
- Some aspects of a subject at the expense of others
- Some students at the expense of others.

He concluded that reliability requires random sampling from the domain of interest and that increasing reliability requires increasing the size of the sample. He suggested that using teacher assessment in certification is attractive as it would increase reliability in relation to test time as well as increasing validity by addressing aspects of construct under-representation. However, he identified problems of a lack of trust ('Fox guarding the henhouse'), problems of biased inferences resulting from construct-irrelevant variance, and the potential introduction of new kinds of construct under-representation.

In his concluding remarks he outlined 'the challenge' – to design an assessment system that is:

- Distributed – so that evidence collection is not undertaken entirely at the end
- Synoptic – so that learning has to accumulate
- Extensive – so that all important aspects are covered (breadth and depth)
- Manageable – so that costs are proportionate to benefits
- Trusted so that stakeholders have faith in the outcomes.

He extended this challenge to delegates whilst recognising the size and complexity of the task as he commented, 'This is not rocket science. It's much harder than that.'

References

Chi, M.T.H., Feltovich, P. & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, **5**, 121–152.

Handy, C. (1994). *The empty raincoat*. London: Hutchinson.

IAEA Conference 2008: papers and presentations. http://www.iaea2008.cambridgeassessment.org.uk/ca/

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, **23**, 2, 13–23.

Salthouse, T. A. (1991). Expertise as the circumvention of human processing limitations. In: K A. Ericsson & J Smith (Eds), *Toward a General Theory of Expertise: Prospects and Limits*. pp.286–300. Cambridge: Cambridge University Press.

Grading examinations using expert judgements from a diverse pool of judges

Nicholas Raikes, Sara Scorey and Hannah Shiell Research Division

This article is based on a paper presented to the 34th annual conference of the International Association for Educational Assessment held in Cambridge, UK, in September 2008. The research was funded jointly by Cambridge Assessment's Research Division and OCR's Operational Research Team. We are grateful to the many examiners, teachers and lecturers who took part.

Introduction

Maintaining grading standards

In this research we investigated a new way of setting grade threshold marks for a General Certificate of Education Advanced Subsidiary (GCE AS) examination. A 'grade threshold mark' defines the boundary between two grades and is the minimum mark required for the higher grade.

GCE examinations are high stakes, content-based assessments often used for university entrance in the United Kingdom. Results are reported as grades, with passing grades from A (top) to E. The examinations are generally held once or twice a year, and on every occasion entirely original question papers are used. The question papers must pass a rigorous quality assurance process, but no formal pre-testing with candidates occurs. Inevitably the question papers vary slightly in difficulty, and so grade threshold marks must be set for each question paper individually, reflecting the particular difficulty level of the paper. With no common questions and no guarantee of common candidates, grade thresholds are set through a process of expert judgement.

We investigated the use of a modified paired comparison technique for equating examinations. We equated an AS biology question paper from June 2007 with one administered in January 2008, and thereby determined the marks on the second question paper which equated to the grade threshold marks previously set for the first. The main focus of the research was whether the results varied with the professional background of the content-experts taking part in the paired comparison exercise.

Paired comparison methods for standard maintaining

Thurstone (1927a, 1927b) introduced methods for constructing an interval scale and simultaneously locating objects on the scale using a process of pairwise comparisons by judges.

A principal advantage of paired comparison methods is that judges make *comparative* judgements, rather than *absolute* judgements. Judges' internal standards cancel out, so that as long as a judge is consistently harsh or lenient, he or she will still make correct *relative ordinal* judgements about the objects in a pair, even if their absolute judgements are wrong. Laming (2004) argues that there is no such thing as absolute judgement, and that all judgements are comparisons of one thing with another and these comparisons are essentially ordinal, adding to the rationale for using paired comparison methods. Simply put, people are better at comparing *concrete* with *concrete* (as in a paired comparison) than *concrete* with *abstract* (as in comparison of an object with an abstract, internal standard).

Examples of the application of Thurstone's paired comparisons method include perceptions of physical properties of objects (e.g. weight), the extremity of attitudes expressed in statements such as statements about capital punishment (Wikipedia, 2008), and the perceived quality of examination scripts. The essential idea is that each object to be judged is successively paired with every other object and the pairs are presented to a number of judges, who work independently. For each pair presented, judges are asked to judge which of the two objects in the pair has more of the attribute being considered. If the objects are reasonably close together, there will be some disagreement. The object judged the 'winner' most frequently is considered to have been perceived to have more of the attribute, and the difference between the objects' numbers of wins is assumed to be related to how far apart the objects were perceived to be in terms of the judged attribute. When all the paired comparisons – that is, the comparisons from each pairing combination and all judges – are considered together, an interval scale can be constructed for the perceived attribute and each object located on the scale using, for example, a Rasch analysis. Bramley (2007) provides a more technical and complete overview, focussed particularly on application of the technique to studies of the comparability of examination standards.

Research aims

The above discussion suggests that a paired comparison methodology might offer an improved basis for inspecting scripts during Awarding. Rather than making absolute judgements about script quality, judges would make relative, ordinal judgements about scripts that were actually in front of them at the time of judgement. This offers the prospect of enabling a wider range and increased number of professionals to be involved in Awarding, since judges would not have to have internalised agreed grade standards. New technology enables digital copies of scripts to be supplied to any number of judges working remotely, so potentially a large number of judges could be involved. Therefore, a paired comparison methodology, coupled with new technology, offers the prospect of more *inclusive* Awarding procedures that take advantage of the professional expertise of a much greater number and range of people. Arguably this would lead to examination standards more clearly grounded in professional communities that the examinations serve. Such large scale paired comparison methods might not need to be employed on every Awarding occasion in order to achieve this end; the full range and number of judges might only need to be consulted periodically, with the smaller Awarding Committee working alone on the intervening occasions.

The aim of the present research was to:

1. Equate two examinations in a GCE assessment unit using a paired comparison method.
2. Compare the scales produced from judgements made by:
   a. senior examiners from the Awarding Committee that recommended the grade boundary marks operationally;
   b. other examiners who marked scripts from the examinations operationally, but did not contribute to Awarding;
   c. teachers who had prepared candidates for the examinations but not marked them;
   d. university lecturers who teach the subject to first year undergraduates (i.e. the university educators who take students on after A Level).
3. Complete and compare the results of the above for two subjects, one assessed primarily with short answer questions and one assessed with essay questions.

The short-answer subject chosen was biology, and the essay subject chosen was sociology. This article reports results for biology only. Work continues on sociology.

Method

Choice of assessment

We used OCR's June 2007 and January 2008 examinations for Advanced Subsidiary GCE Biology Unit 2801, Biology Foundation[1]. We chose this unit because it had a relatively high entry in both January and June and was assessed using a range of item types, including single word answers, calculations, short answers of one or two sentences and more extended answers of up to around an A4 page of factual writing. Both examinations were marked out of 60 raw marks and candidates were allowed one hour.

Grade boundaries had been set operationally for both of these examinations. The equating exercise conducted for the research was for research purposes only. We imagined that the June 2007 boundary marks were known (as indeed they were) and that we were trying to carry forward the grading standards and set boundary marks for the January 2008 examination.

Scripts

We used real scripts from the live examinations in the range 14–52 (out of 60) raw marks.

Seven scripts on each total raw mark were chosen at random from each examination (only six scripts were available on some marks, and in these cases all available scripts were chosen). The chosen scripts were obtained from Cambridge Assessment's warehouse and the item marks keyed. The marks were analysed using a separate Rasch partial credit model for each examination and the best fitting script on each mark in the range 14–52 was selected for use in the study. In this way we tried to ensure that the scripts used were reasonably typical of those on each mark.

The selected scripts were scanned and the marks, examiner annotations and all candidate and centre details deleted from the resulting images. It is necessary to delete marks from the scripts seen by judges making paired comparisons since otherwise the comparisons are likely to be largely based on a comparison of the marks rather than of perceived quality. Scripts were allocated an identification number at random and the identifier was written at the top of page 1 of each script. Multiple copies of the 'clean' images were printed for use in the study – we decided to send participants hard copies, rather than electronic copies for on-screen viewing, so that we could control the judges' experience as much as possible and thereby minimise the risk of introducing extraneous variables into the research.

Participants

The following numbers of participants were recruited:

Members of the current Awarding Committee	6
Examiners	48
Teachers	57
University lecturers	54

We paid participants for their time: 2 hours per person for the examiners, teachers and lecturers; 16 hours per person for members of the Awarding Committee (this group was much smaller than the others, so each person had to make more comparisons so that overall the groups made an approximately equal number of comparisons). The paid time was intended to cover all participants' activities, that is, preparation and feedback as well as performing the rankings.

Paired comparison method

We used Bramley's (2005, 2007) rank ordering method to generate inferred paired comparisons. Script copies were sent to judges in packs of three – we chose threes because we judged that this enabled us to make efficient use of our judges' time whilst keeping the task for judges plausibly achievable, that is, to sort the scripts, on the basis of an holistic judgement, into best, middle and worst. Black (2008) reports successful use of packs of three scripts, and Bramley *et al*. (2008) provide evidence for the validity of the rank ordering method.

Triples design

We had 39 scripts from each examination, one on each raw mark in the range 14–52 inclusive, giving 78 scripts in total. A total of 3,081 different pairs can be constructed from these 78 scripts.

We estimated that it would take participants 10–15 minutes to rank-order a pack of three scripts, depending on the particular scripts in the pack and a participant's speed of working. We decided to ask members of the Awarding Committee to rank-order 60 packs each, and the other participants 8 packs each. The Awarders would therefore complete the smallest number of packs (6 judges × 60 packs each = 360 packs). Even so, since we infer 3 paired comparisons per pack, this would enable the Awarders to judge around a third of the 3,081 possible pairs; with the addition of a restriction to avoid using pairs where scripts are more than a third of the 60 available marks apart, coverage is adequate. The restricted range is reasonable since it is not plausible that the two examinations' difficulties could be so poorly aligned that an adjustment of as much as 20 marks would be required to equate them.

A total of 400 triples were designed as follows:

- Each script was required to appear in an approximately equal number of triples (15 or 16, i.e. 400 triples × 3 script-copies divided by 78 scripts = 15.4 triples per script).

1 Candidates must take a total of three units for an AS qualification in biology, with a further three at the more demanding A2 level for a full Advanced GCE qualification in biology.

- No particular script pairing was allowed to appear in more than two triples.
- Each triple was required to contain scripts from both examinations. Half the triples contained a single June 2007 script and two January 2008 scripts, the other half contained two June 2007 scripts and a single January 2008 script.
- Every script appeared as the 'single' script in an approximately equal number of triples.
- When the scripts in a triple were ordered by raw mark[2], the number of triples where the 'single' script was top was required to be approximately equal to the number of triples where it was middle and the number where it was bottom. This was to ensure that judges didn't come to expect the single script always to occupy the same position.
- The range of raw marks spanned by a triple was required to be no more than 20 (one third of the maximum raw mark available for the assessment).

Triple allocation

The 400 triples were sorted into a random order, given a sequential identification number and allocated to each group of participants in that order. The first 60 triples were allocated to the first Awarder, the next 60 to the second Awarder, and so on until all 6 Awarders had been allocated their 60 triples (the final 40 triples were not allocated to Awarders). Allocations were repeated for the other groups of participants, but this time only eight triples were allocated per person – that is, the first 8 triples were allocated to the first examiner, teacher and lecturer, the next 8 to the second examiner, teacher and lecturer, and so on. More than 50 teachers and 50 lecturers took part, so more than 400 triples were required – for these two groups, the 51st participant received the same triples as the first participant, the 52nd the same as the second, and so on until every judge had been allocated 8 triples.

Materials supplied to participants

Script packs were constructed in accordance with the above triple allocations, with each triple having its own pack. Participants were sent:

- their script packs;
- cut-down mark schemes containing illustrative correct answers for every question;
- machine-readable record sheets for recording their rank order decisions;
- a short feedback questionnaire.

Participants were instructed to work through their packs in the order of the pack identifiers. The instructions required participants to:

> *Place the three scripts in each pack into a single rank order from best to worst, based on the quality of the candidates' answers. You may use any method you wish to do this, based on scanning the scripts and using your own judgement to summarise their relative merits, but you must not re-mark the scripts. You should endeavour to make an holistic judgement about each script's quality.* ***Remember, this is not a re-marking exercise****.*
>
> *No tied ranks are allowed. … Do not agonise for ages over the correct rank order if scripts appear to be of exactly the same standard; several judges will see the scripts and we will infer that scripts are of equal standard when judges are split approximately 50–50 on their relative standard.*

Scale construction and script location

The ranking data were converted to inferred paired comparison data (for example, if a judge put three scripts into the order script-2 (top), script-1, script-3, then the inferred paired comparisons were: script-2 beats script-1, script-2 beats script-3 and script-1 beats script-3). Each group's paired comparison data were analysed separately using a Rasch model to construct the scale and estimate the location (measure) of each sample script on this scale (Andrich, 1978). FACETS software was used to estimate the parameters (Linacre, 2006).

Results

Intra-group reliability

Table 1 presents internal reliability data for the scales and script-measures produced from each group's comparisons. The reliability coefficient reported is the Rasch equivalent of Cronbach's alpha, and the figures indicate very high and similar reliabilities for all four groups of judges. The correlations between the operational raw marks and the measures produced in the research are also very high for all four groups for both examinations. It is worth reflecting that we would not expect to get exactly the same marks if we had the scripts re-marked, so the correlations are very impressive. The last column in Table 1 gives the percentage of paired comparison results made by each group that were

2 Raw marks were removed from the script copies seen by judges, but the researchers kept a record of the live raw marks given to each script.

Table 1: Internal reliability data for the scales and measures produced from each group's comparisons

	Judges	*Triples*	*Pairs*	*Reliability**	*Correlation between raw mark & measure*		*Paired comparisons consistent with measures*
					June	*January*	
Awarders	6	359	1077	0.95	0.95	0.91	81%
Examiners	48	383	1149	0.97	0.96	0.95	84%
Teachers	57	455	1365	0.97	0.95	0.95	83%
Lecturers	54	431	1293	0.96	0.93	0.93	82%

* Separation reliability

consistent with the script-measures estimated from that group's rankings. This is an indicator of the level of agreement between the judges in a group, and the similar figures indicate similar levels of inter-judge agreement for each group.

Inter-group reliability

Table 2 gives the correlation among the script-measures estimated from each group's rankings. The correlations are all high and similar to each other, indicating a high degree of inter-group reliability.

Table 2: Correlation matrix for the script-measures estimated from each group's rankings

	Awarders	*Examiners*	*Teachers*	*Lecturers*
Awarders	1.00	0.93	0.94	0.92
Examiners	0.93	1.00	0.95	0.95
Teachers	0.94	0.95	1.00	0.94
Lecturers	0.92	0.95	0.94	1.00

Estimated grade boundaries for January 2008

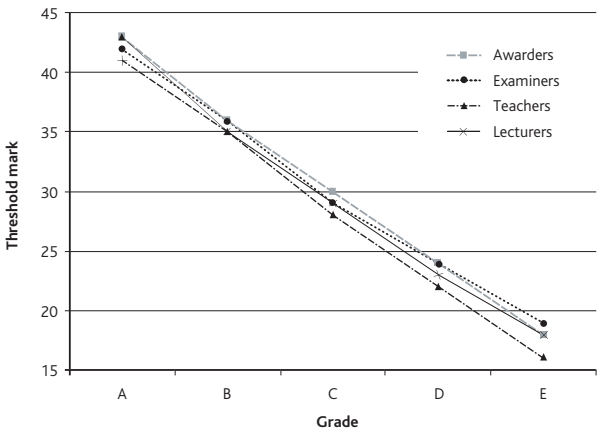
Table 3 gives the grade boundary marks estimated from each group's rankings for the January 2008 examination. Figure 1 presents the same information graphically (the lines between the points have been drawn in for clarity but have no meaning). The figures are similar for each group, with a maximum spread of 3 marks (for the E boundary). The boundaries estimated from the Awarders, examiners and lecturers' data are all within just 1 mark of each other for grades B–E. To place this in context, when an Awarding Committee inspects scripts operationally using the top-down, bottom-up procedure described in the introduction, the gap between the upper and lower limiting marks for a key boundary (i.e. the range in which the key boundary is expected to lie) is typically between 2 and 5 marks' wide for A Level science units. There was a remarkable degree of agreement between the boundaries estimated from each group's ranking data in the present study.

The teachers' data yielded the lowest estimates for the boundaries at C–E. Although it is tempting to conclude from this that the teachers were more generous than the other groups at these grades, the corollary is that they judged the June 2007 scripts slightly more harshly than the other groups.

Table 3: Grade boundary marks estimated from each group's rankings for the January 2008 examination

	Minimum mark required for grade				
	A	*B*	*C*	*D*	*E*
Awarders	42	36	29	24	19
Examiners	43	36	30	24	18
Teachers	43	35	28	22	16
Lecturers	41	35	29	23	18

Figure 1: Grade boundary marks estimated from each group's rankings for the January 2008 examination



Conclusion

In this study we investigated the potential of an adapted Thurstone paired comparisons methodology for enabling a greater range and number of educational professionals to contribute to decisions about where grade boundaries should be located on examinations.

The research was done using an OCR GCE AS biology assessment, though the results should be applicable to similar examinations. Examinations administered in June 2007 and January 2008 were equated in the study using paired comparison data from the following four groups of judges:

- Senior examiners from the Awarding Committee that recommended the grade boundary marks operationally.
- Other examiners who marked scripts from the examinations operationally, but did not contribute to Awarding.
- Teachers that had prepared candidates for the examinations but not marked them.
- University lecturers who taught the subject to first year undergraduates.

Each group's paired comparison data were analysed separately using a Rasch model to construct a singe interval scale for both examinations and to estimate the location (measure) of each sample script on this scale.

We found very high levels of intra-group and inter-group reliability for the scales and measures estimated from all four groups' judgements. When boundary marks for January 2008 were estimated, there was considerable agreement between the estimates made from each group's data. Indeed for four of the boundaries (grades B, C, D and E), the estimates from the Awarders', examiners' and lecturers' data were no more than 1 mark apart, and none of the estimates were more than 3 marks apart.

We conclude from these findings that the examiners, teachers, lecturers and members of the current Awarding Committee made very similar judgements. If live Awarding procedures were changed so as to include a paired comparisons exercise, examiners, teachers and lecturers could take part without compromising reliability.

The next phase of the current research is to analyse feedback from participants and to repeat the entire analyses with similar data collected in the context of AS GCE sociology, which is assessed via essay questions.

We envisage that large scale paired comparison exercises conducted as part of operational Awarding would be done using digital copies of scripts viewed by judges on screen, rather than the hard copies used in the present research. We recommend that further research or trials be conducted to investigate whether judges make similar judgements when viewing scripts on screen as on paper. We also recommend that research be conducted to investigate whether other groups of stakeholders – subject experts from industry, for example – make judgements consistent

with those of judges from the education sector, with the aim of also including representatives from these further stakeholder groups in Awarding.

References

Andrich, D. (1978). Relationships between the Thurstone and Rasch approaches to item scaling. *Applied Psychological Measurement*, **2**, 449–460.

Black, B. (2008). *Using an adapted rank-ordering method to investigate January versus June awarding standards*. Paper presented at the Fourth Biennial EARLI/ Northumbria Assessment Conference, Berlin, Germany, August 2008.

Bramley, T. (2005). A rank-ordering method for equating tests by expert judgment. *Journal of Applied Measurement*, **6**, 2, 202–223.

Bramley, T. (2007). Paired comparison methods. In: P. Newton, J-A Baird, H. Goldstein, H. Patrick and P. Tymms (Eds.), *Techniques for monitoring the comparability of examination standards* (chapter 7). London: QCA.

Bramley, T., Gill, T. and Black, B. (2008). *Evaluating the rank-ordering method for standard maintaining*. Paper presented at the 34th Annual Conference of the International Association for Educational Assessment, Cambridge, UK, September 2008.

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

Linacre, J.M. (2006). *FACETS [Computer program, version 3.60.0]*. www.winsteps.com

Thurstone, L.L. (1927a). Psychophysical analysis. *American Journal of Psychology*, **38**, 368–389. In: Thurstone, L.L. (1959). *The measurement of values* (chapter 2). Chicago, Illinois: University of Chicago Press.

Thurstone, L.L. (1927b). A law of comparative judgment. *Psychological Review*, **34**, 273–286. In: Thurstone, L.L. (1959). *The measurement of values* (chapter 3). Chicago, Illinois: University of Chicago Press.

Wikipedia (2008). *Law of comparative judgment*. http://en.wikipedia.org/wiki/Law_of_comparative_judgment Accessed 11 July 2008.

Using 'thinking aloud' to investigate judgements about A-level standards: Does verbalising thoughts result in different decisions?

Dr Jackie Greatorex and Rita Nádas Research Division

This article is based on a paper presented at the British Educational Research Association Annual Conference, September 2008, Edinburgh.

Abstract

Background

The 'think aloud' method entails people verbalising their thoughts while they do tasks, resulting in 'verbal protocols'. The verbal protocols are analysed by researchers to identify the cognitive strategies and processes as well as the factors that affect decision making. Verbal protocols have been widely used to study decisions in educational assessment. The main methodological concern about using verbal protocols is whether thinking aloud compromises ecological validity (the authenticity of the thought processes) and thus the decision outcomes. Researchers have investigated to what extent verbalising affected the thinking processes under investigation in a variety of settings. Currently, the research literature generally is inconclusive; most results show just longer performance times and no alternative task outcome.

Previous research on *marking* collected decision outcomes from two conditions:

1. marking silently;
2. marking whilst thinking aloud.

The mark to re-mark differences were the same in the two conditions. However, it is important to confirm whether verbalising affects decisions about grading standards. Therefore, our main aim was to compare the outcomes of senior examiners making decisions about *grading* standards silently as opposed to whilst thinking aloud. Our article draws from a wider project taking three approaches to grading.

Method

In experimental conditions senior examiners made decisions about A-level grading standards for a science examination both silently and whilst thinking aloud. Three approaches to grading were used in the experiment. All scripts included in the research had achieved a grade A or B in the live[1] examination. The decisions from the silent and verbalising conditions were statistically compared.

Findings

Our interim findings suggest that verbalising made little difference to the participants' decisions; this is in line with previous research in other contexts. The findings reassure us that the verbal protocols are a useful method for research about decision making in both marking and grading.

Background

The 'think aloud' method entails people verbalising their thoughts while they perform tasks. The resulting 'verbal protocols' are then analysed by researchers. The think aloud procedure is an established method of researching what people pay attention to, or what cognitive strategies they are using when they do various complex tasks (e.g. Van Someren

1 Live is used to denote the examination or procedures taking place 'for real' rather than as part of an experimental setting.

et al., 1994; Taylor and Dionne, 2000). Verbal protocols have been widely used to investigate decision making processes in educational assessment (Cumming, 1990; Sanderson, 2001). Various studies carried out by Cambridge Assessment used verbal protocols to investigate the judgement process involved in marking varied A-level and GCSE examinations (Suto and Greatorex, 2008a and b; Crisp, 2007 and 2008a), as well as to explore the process of judging grading standards in A-levels (Crisp, 2007 and 2008b). One frequent question to Crisp, Suto and Greatorex from researchers and assessment professionals was whether the method of verbalising alters the outcomes of decision processes. Researchers have studied to what extent verbalising affected the cognitive processes in various settings (e.g. Ummelen and Neutelings, 2000). Although the research is currently inconclusive most results show longer performance times and no alternative task outcome (Krahmer and Ummelen, 2004).

There is one piece of research which answers our question in the context of A-level *marking*. Crisp (2008c) collected decision outcomes from two conditions:

1. marking silently;
2. marking whilst thinking aloud.

The mark to re-mark differences in the two conditions were similar. However, it is important to confirm whether verbalising affects decisions about *grading* standards because marking and grading are two distinct but linked procedures in the context of A-level and GCSE examinations. Grading (awarding) meetings involve senior examiners recommending grade boundaries after marking has been completed.

In this article we aim to answer the frequently asked question of whether thinking aloud results in different decisions about A-level grading standards. Ensuring the robustness of research on the psychology of decision making processes in assessment is of crucial importance, especially when using the think aloud method. Therefore, our article draws from a wider project where the main aim was to find out more about cognitive decision making processes used to make judgements about grading standards. As well as studying the decision making processes, the outcomes of the decisions were also considered to be important.

In the wider project five senior examiners[2] made decisions about A-level grading standards for a science examination both silently and whilst thinking aloud. All the decisions were made in experimental conditions for research purposes. Three approaches are considered:

i. awarding – part of the conventional approach to recommending grade boundaries,
ii. Thurstone pairs,
iii. rank ordering.

The latter two were suggested as possible future methods of recommending grade boundaries by Pollitt and Elliott (2003a and b), and Black and Bramley (2008). They have also been used in a series of comparability studies (e.g. Forster and Gray, 2000; Arlett, 2003; Greatorex *et al.*, 2002, 2003; Edwards and Adams, 2002, 2003; Guthrie, 2003; Bramley *et al.*, 1998; Townley, 2007; Black and Bramley, 2008). The focus of this article will be a statistical comparison of the decisions from the silent and verbalising conditions, and is the first in a series of linked studies which make up the wider project. So far only one other study from the wider project is complete, in which Greatorex *et al.* (2008) analysed which items the participants attended to whilst making decisions and whether these items were likely to be helpful in decision making.

In this research we focus on one decision-making phase of awarding which involves the awarding committee judging whether a small number of examples of candidates' work on particular marks show the distinguishing characteristics of performance at a particular grade. For a fuller description see Cresswell (1997), QCA (2008) or Greatorex (2003). The candidates' work is usually examination scripts but might be a recording of a drama or musical performance or an artefact such as a painting. Thurstone pairs and rank ordering, as well as examples of their use in comparability studies, have been frequently described in the literature; see for example Bramley *et al.* (1998), Arlett (2003), Greatorex *et al.* (2002, 2003), Edwards and Adams (2002, 2003), Guthrie (2003) and Townley (2007). Therefore, we will only provide a summary here. Thurstone pairs and rank ordering involve a group of experts judging the quality of candidates' work. In Thurstone pairs in this context each expert compares a pair of scripts, with each pair constituting a script from the live examination and the archive examination. Each expert decides which of two scripts shows evidence of better candidate performance, without re-marking the scripts. This is repeated for a variety of pairs of scripts. When all the necessary comparisons have been made, they are statistically analysed (using Rasch). The results of the analysis can be used to identify a small range of marks within which the live boundary should lie for the standard from last year to be maintained. In rank ordering each expert receives small samples of live and archive scripts which they rank according to the candidates' performance. This is repeated for a number of overlapping samples of scripts. The outcomes of the rankings are submitted to the same statistical analysis as above. Again the statistics can be used to identify a small range of marks within which the live boundary should lie.

There are a number of aspects of awarding meetings and scripts that positively and negatively influence judgements of gradeworthiness (Cresswell, 1997; Murphy *et al.*, 1995; Crisp, 2007; Baird, 2000; Baird and Scharaschkin, 2002; Scharaschkin and Baird, 2000). To understand some of the resulting difficulties we have to bear in mind that A-level and GCSE examinations have a principle of compensation, according to which candidates gain marks for their strengths, and there is more than one way to achieve a grade. Arguably, one issue influencing examiners' grading decisions is that sometimes the visibility of marks given to candidates' responses and the marks available on the question paper become extraneous information. Two conundrums relate to the principle of compensation and the visibility of marks on scripts:

- Some awarding committee members pay particular attention to questions and marks which are believed to differentiate between performances at particular grades (Murphy *et al.*, 1995; Greatorex *et al.*, 2008). This belief might be well or ill founded (Murphy *et al.*, 1995). Focussing on particular questions at the expense of other questions is not aligned with the principle of compensation[3].

2 All participants in the wider project had been involved in the live Award for the examination and met the criteria used in many comparability studies for recruiting participants.

3 Grade descriptors are a written summary of the features of performance at particular grades. It is important to note these are indicators of typical performance and not criteria to be met. The grade descriptors are cues to memory which can be used in Awarding meetings. Some might argue that GCSEs or A-levels did not have a principle of compensation because some grade descriptors, including those for the science examination in the research, refer to some high grade performances as being consistently high achieving. However, the principle of compensation holds as these are indicators not criteria, so students do not have to have all the characteristics in the grade descriptors to get the grade. For more details about grade descriptors see Greatorex (2001, 2002, 2003b).

Psychological research from a variety of contexts presented by Greatorex (2007) and later Greatorex *et al*. (2008) suggests that humans are not particularly good at combining information to make decisions. Therefore, focussing judgements on particular questions might be a successful approach to decision making, if the questions are a good proxy for the whole of the examination. After all, the alternative strategy – judgements about whole scripts – involves mentally combining an examinee's answers to all questions in the examination.

- It has been established that the consistency of candidates' performance across questions on an examination paper influences the severity of judgements of gradeworthiness (Cresswell, 1997; Scharaschkin and Baird, 2000). Again, this is not aligned with the principle of compensation.

Given that, arguably, marks can act as extraneous information and that scripts are cleaned of marks in some comparability studies, we decided to include the visibility of marks as a variable in our research.

How reliable are the judgements made using each method?

This research also provided an opportunity to compare the reliability of judgements made under a variety of conditions by using the different methods mentioned above and by adding the visibility of marks as a variable. The reliability of awarding judgements is a well researched topic. As has been apparent for some time, the precision of awarding is less than perfect. For example, see Willmott and Nuttall's (1975) work about the General Certificate of Education O Levels and the Certificate of Secondary Education, the predecessor qualifications of GCSE. In later research, Good and Cresswell (1988) replicated some awarding meetings for French, History and Physics. Good and Cresswell (1988; p. 23 in Cresswell, 2000) concluded that 'different groups of grade Awarders can reach decisions about final grade boundaries which are sufficiently similar to be acceptable, given the inherent imprecision of the examining process'. They aimed to find out what percentage of candidates' grades would have changed if one awarding team's judgements were substituted for another's. They found that 13% of candidates' grades would have changed in French, 17% in Physics and 38% in History. This finding might raise questions about the reliability of awarding procedure; however, Cresswell (2000) cites Willmut (1981) who showed that the change of 38% of grades outcomes corresponds approximately to an inter-examiner reliability coefficient of 0.96, which is generally considered to be of very high inter-rater reliability. Previous research on awarding has established that the severity of judgements of gradeworthiness can be influenced by several factors of arguably varying validity. For instance:

i. the archive scripts provided (Baird, 2000),
ii. the consistency of performance in scripts (Scharaschkin and Baird, 2000) and
iii. whether the examiners see candidates' work from one examination or the work of candidates from the whole qualification (Baird and Scharaschkin, 2002).

In summary, experiments suggest that awarding committees do not make perfectly consistent decisions.

There is little research in the public domain comparing the reliability of decisions made in studies using Thurstone pairs and rank ordering with other approaches to assessment. The main body of evidence in the public domain is by Kimbell *et al*. (2007), who used a mixture of Thurstone pairs and rank ordering to assess a series of Design and Technology portfolios in a pilot study. They claim that their approach to assessment is highly reliable (Kimbell *et al*., 2007). Of course, there is already a large literature illustrating that the reliability of marking is generally less than perfect, for example, see Hartog and Rhodes (1935), Pillner (1968), Willmott and Nuttall (1975), Newton (1996), Pinot de Moira *et al*. (2002), Raikes and Massey (2007) and Vidal Rodeiro (2007). Overall the research is inconclusive regarding whether Thurstone pairs/rank ordering decisions are of similar reliability to more conventional approaches to making assessment decisions. The present research adds to the accumulation of evidence.

Method

Verbal protocols result from participants verbalising their thoughts as, or after, they perform a complex cognitive activity. This is an established method of studying what people pay attention to, or what strategies they use when they are undertaking a variety of complex cognitive tasks (e.g. Van Someren *et al*., 1994; Taylor and Dionne, 2000), including decisions in educational assessment (Crisp, 2008a and b; Suto and Greatorex, 2008a and b; Green 1998; Cumming, 1990; Vaughan, 1992; Weigle, 1994; Milanovic *et al*., 1996).

One of the most established approaches to using verbal protocols is explained by Ericsson and Simon (1993). The approach is also recommended by Krahmer and Ummelen (2004) because it has a sound theoretical basis underpinning it, which some rival approaches do not. The thinking aloud procedure in this research reflected Ericsson and Simon's principles. For instance, the participants had a practice session, and were not interrupted whilst providing the 'real' verbal protocol. The exception to this principle was when a participant was silent for some time and the researcher said 'please keep talking'. The participants were asked to say which script and item they were looking at in the verbal protocols to facilitate the analysis.

Initially, the participants made awarding, Thurstone pairs and rank ordering judgements silently at home. The tasks were then repeated whilst thinking aloud as the main data collection phase. There were some differences between the script samples and procedures for the decisions made silently and whilst thinking aloud (more details are given later). This was because there were only a limited number of scripts to work with and the arrangements for the main data collection phase took precedence over the arrangements for the decisions made silently.

Examination

An AS-level science examination from 2005 and another from 2006 were used in the research. The examinations were from the same qualification and specification. The candidates' work is likely to provide evidence of numerical skills, written skills, use of diagrams and knowledge and understanding. Therefore, research results from this examination might be more generalisable than those from a different subject.

For each examination a total of 45 marks were available. In the live examination the question papers were given to candidates as a form in which the items and source material (e.g. diagrams) were presented along with an answer space into which they added their responses. All the items were worth between 1 and 6 marks. Additionally, one mark

was available on each question paper for QWC (quality of written communication) and this was associated with one item on each paper worth 6 marks (plus 1 mark for QWC). The mark scheme was a points based mark scheme.

Script samples

The script samples constituted scripts with total marks within the range of marks considered in the recommendation for the grade A boundary in the awarding meeting (33 to 37 for 2005 and 28 to 34 for 2006). The live grade A boundary was 35 marks for the 2005 examination and 31 marks for the 2006 examination. The frequency of scripts in the sample for the decisions made whilst thinking aloud and the decisions made silently are given in Table 1.

Table 1: Frequency of the scripts with a particular mark

Total marks from 2005	*frequency of scripts in the silent conditions*	*frequency of scripts in the thinking aloud conditions*	***Total marks from 2006***	*frequency of scripts in the silent conditions*	*frequency of scripts in the thinking aloud conditions*
33	2	3	**28**	0	3
34	2	3	**29**	0	5
35	2	7	**30**	2	5
36	2	3	**31**	2	3
37	2	3	**32**	2	4
			33	2	5
			34	2	4
Total	**10**	**19**		**10**	**29**

Participants

Five senior examiners who were involved in recommending live grade boundaries for the AS-level examination in either 2005 and/or 2006 took part in the research.

Conditions

The awarding conditions reflected the aspect of awarding where individual awarding committee members evaluate scripts, prior to coming to a collective view about where the grade boundary should be. The rank ordering and Thurstone pairs conditions were intended to reflect current/best practices in previous studies. For all conditions some minor adjustments were made to current/best practices for the purposes of this research (e.g. asking participants to think aloud).

In our study photocopies of the scripts were used rather than the original scripts. For each method the scripts were presented as they are normally presented: awarding with marks visible, Thurstone pairs with marks visible[4] and rank ordering with scripts cleaned of marks. For awarding and Thurstone pairs the procedures were also undertaken with the scripts cleaned of marks. This experimental control was introduced given the arguably extraneous influence of visible marks in some awarding judgements (Murphy *et al*., 1995; Cresswell, 1997; Scharaschkin and Baird, 2000).

This gave us 10 different experimental conditions:

Table 2: Experimental conditions

Awarding method	*Scripts cleaned of marks (clean) or with marks visible (visible)*	*Decisions made silently (silent) or whilst thinking aloud (VP)*	*Term to be used to refer to the condition*
awarding	Clean	silent	awarding clean silent
awarding	Clean	VP	awarding clean VP
awarding	Visible	silent	awarding visible silent
awarding	Visible	VP	awarding visible VP
rank ordering	Clean	silent	rank ordering clean silent
rank ordering	Clean	VP	rank ordering clean VP
Thurstone pairs	Clean	silent	Thurstone pairs clean silent
Thurstone pairs	Clean	VP	Thurstone pairs clean VP
Thurstone pairs	Visible	silent	Thurstone pairs visible silent
Thurstone pairs	Visible	VP	Thurstone pairs visible VP

Guarding against order effects

Scripts were included in more than one condition when the decisions were made silently. However, each participant undertook the tasks in a different order to guard against order effects[5].

For the main data collection phase each participant experienced the conditions one after the other in the Cambridge offices with a researcher present. (Unfortunately, sometimes the participants did not complete all the tasks due to time constraints and therefore there were some missing data). Three precautions were followed to minimise order effects:

- each participant experienced each condition in a particular order;
- in between undertaking one verbal protocol condition and the next the participants took a break or undertook a distractor task. In the distractor task the participants considered some examination questions from an international syllabus (in the same school subject) at a lower level and rated how accurately they thought different groups of examiners would mark the questions;
- for any given participant each script only appeared in one condition; this was also to guard against participants remembering the scripts.

Additionally, the scripts were designated across tasks and participants to avoid interactions.

For all the conditions, the instructions used for making decisions silently were similar to those used in the main data collection phase. For all conditions, only the question paper, scripts and mark scheme were available for reference. The participants were asked not to use the mark scheme to re-mark the scripts. (The additional information that is provided in live awarding meetings was not provided in this research as it might have influenced the judgements in the other conditions).

For the main data collection phase the instructions used in the Thurstone pairs and rank ordering conditions closely resembled the instructions from the most recent Cambridge Assessment studies in the

4 Scripts with marks visible have been used in most of the recent inter-Awarding Body comparability studies for UK examinations. These studies were conducted using Thurstone pairs. To explain why cleaning scripts of marks was not necessary in these studies we need to consider what the participants were doing. The participants were asked to make comparisons at the qualification rather than the examination level. That is they would be comparing say three scripts from one candidate who took AQA A-level Chemistry and another three scripts from another candidate who took OCR A-level Chemistry. For a participant to work out a candidate's overall qualification mark they would need to take into account the proportion of the available marks achieved, uniform mark scale calculations as appropriate, the weighting applied to each examination to provide the final overall qualification grade and so on. Given this complexity it is arguably harder to work out how to compare the performances based solely on which candidate has achieved the higher proportion of marks than to make a qualitative judgement about the quality of the performance.

5 Order effects in this research could be the order in which the conditions and/or scripts were experienced and thereby affecting the decisions.

public domain (Pollitt and Crisp, 2004; Black and Bramley, 2008) with any necessary changes in details for the purposes of this study (e.g. thinking aloud). This was to ensure that current/best practices were followed and to ensure the instructions matched those generally used in studies as deviations from usual practices which would invalidate the research.

The verbal protocols were digitally recorded with the permission of the participants. Subsequently, the digitally recorded information was transcribed.

Analysis

For each awarding condition, the proportion of occasions on which 2006 scripts with a particular mark were judged worthy of a higher grade was calculated. For Thurstone pairs and rank ordering, we calculated the proportion of occasions on which 2005 scripts on a particular mark were judged to be better than scripts from 2006 on a particular mark. For instance, we calculated the proportion of occasions on which 2005 scripts with a mark of 35 were judged to be better than (winning against[6]) 2006 scripts of 29 marks. The calculations were undertaken for each mark from each year. The figures were calculated separately for the scripts with marks visible and the scripts cleaned of marks. (Note that for rank ordering the experts ranked two samples of scripts and the outcomes could not be statistically combined so the results have been presented separately.)

The resulting patterns from the decisions made silently and whilst thinking aloud were compared by scanning the figures. The figures are indicated in Tables 3 to 5. We predicted we would gain an approximate increasing monotonic relationship from the bottom left corner to the top right corner in each table because we expected that:

- In the awarding conditions 2006 scripts with higher marks will be judged worthy of a higher grade on a higher proportion of occasions. For example, 28-mark scripts (from 2006) should be judged worthy of grade A on a smaller proportion of occasions than 37-mark scripts (from 2006).
- In the Thurstone pairs and rank ordering conditions the proportion of occasions on which 2005 scripts are judged better than 2006 scripts will increase as the 2005 total mark increases. For example, in any comparison 33-mark scripts (from 2005) should win against 28-mark scripts (from 2006) on a smaller proportion of occasions than 37-mark scripts (from 2005) should win against 28-mark scripts (from 2006). Also, in any comparison 33-mark scripts (from 2005) should win against 34-mark scripts (from 2006) on a smaller

6 When a script is judged to be better in quality than another we sometimes refer to this as 'winning against' another script.

Table 3: The proportion of occasions on which 2006 scripts with a particular mark were judged worthy of grade A in the awarding conditions

Silent

Proportion of occasions on which 2006 scripts were judged worthy of grade A	2006 Mark	30	31	32	33	34
	1					**v**
	0.9					
	0.8				**v**	
	0.75			**v**		
	0.6	**c**	**c**			
	0.5					
	0.4				**c**	**c**
	0.3					
	0.25		**v**	**c**		
	0.1					
	0	**v**				

Thinking Aloud

Proportion of occasions on which 2006 scripts were judged worthy of grade A	2006 Mark	28	29	30	31	32	33	34
	1					**c**		**v**
	0.9							
	0.8						**v**	
	0.75				**c**			**c**
	0.6					**v**		
	0.5							
	0.4							
	0.3				**v**			
	0.25		**c**	**c**			**c**	
	0.1							
	0	**v, c**	**v**	**v**				

c = Cleaned **v** = Visible All marks in the grid are from 2006

How to use Table 3

As an example of how to use Table 3 we can see that the proportion of occasions that 32-mark scripts from 2006 (found in the top row) were judged worthy of grade A for decisions made silently was 0.25 (found in the left hand column) for the cleaned of marks condition (indicated by a **c** in the grid) and 0.75 for the marks visible condition (indicated by a **v** in the grid).

In Table 3 we would expect to get an approximate increasing monotonic relationship from the bottom left hand corner to the top right hand corner. This is because the higher the 2006 mark the greater the proportion of occasions on which scripts should be judged worthy of grade A, irrespective of the visibility of their marks.

proportion of occasions than 37-mark scripts (from 2005) should win against 34-mark scripts (from 2006).

All the marks shown in the analysis refer to the total marks achieved by candidates on the examination paper in question.

According to some awarding committee members it is difficult to make decisions about 'rogue' or apparently atypical scripts, and they usually avoid using such scripts in making recommendations for grade boundaries in live awarding meetings, for example, Murphy *et al*. (1995). However, it was assumed for this study that all scripts in the sample on a particular mark had the characteristics of performance at that mark. This is a reasonable assumption given that in live contexts all scripts on a particular mark are awarded a particular grade.

Results

See Tables 3, 4 and 5.

Findings given in Table 3

Broadly speaking there is little difference between the pattern of decisions made silently in comparison with the pattern of decisions made whilst thinking aloud. This reinforces the findings in previous literature.

However, there was a considerable difference between the pattern of decisions made with scripts with marks visible and the pattern of decisions made on scripts cleaned of marks. When the marks were visible the expected pattern was evident. We can see that the expected pattern was not anywhere near as clear for decisions made when the scripts were cleaned of marks.

Findings given in Table 4

Overall, in general, the expected pattern is evident, for decisions made:

i. silently and whilst thinking aloud,
ii. with scripts with marks visible and scripts cleaned of marks.

However, there are some scripts which do not conform to the pattern.

Findings given in Table 5

For both decisions made silently and whilst thinking aloud the broad pattern is similar; in both cases as the 2005 marks increase the proportion of occasions on which the 2005 scripts win also increases. However, there are a few scripts which seemed to be ranked lower or higher than would be expected given the mark achieved by the candidate.

Table 4: The proportion of occasions on which 2005 scripts won against 2006 scripts in the Thurstone pairs conditions

Silent

Proportion of occasions on which 2005 scripts won against 2006 scripts

2005 Mark	*33*	*34*	*35*	*36*	*37*
1				**30v**	**34c, 31v**
0.8		**30v, 32v**	**32v**	**32c, 31v**	**34v**
		30c	**31c**		
0.6	**34v**		**33v**	**34c**	**33c**
0.4	**32c, 33c**				
	33v				
0.2					
0					

Thinking Aloud

Proportion of occasions on which 2006 scripts were judged worthy of grade A

2005 Mark	*33*	*34*	*35*	*36*	*37*
1	**29v, 32v**	**28c, 33v, 34v**	**29c, 32c**	**30c, 29v, 30v**	**28c, 31c**
		30v	**29v**		**31v, 32v, 33v**
0.8					
0.6					
	30c, 30v	**34c**	**34v**		**33c**
0.4					
0.2					
	31v				
0	**29c, 32c, 34c**	**29c, 31c**	**33v**	**29c, 33c**	**32c**

c = Cleaned **v** = Visible

How to use Table 4

As an example of how to use Table 4 in the marks visible condition the proportion of occasions on which 2005 scripts with 34 marks (found in the top row) won against 2006 scripts with 30 marks (found in the grid) was 0.8 (found in the left hand column), also the proportion of occasions on which 2005 scripts with 34 marks won against 2006 scripts with 32 marks was 0.8.

In Table 4 we expect to see an approximate increasing monotonic relationship from the bottom left hand corner to the top right hand corner. This is because the proportion of occasions on which 2005 scripts should win against 2006 scripts should increase as the 2005 marks increase. The higher the 2005 marks, the larger the proportion of occasions on which 2005 scripts should win against 2006 scripts, irrespective of the 2006 script mark. For example, in any comparison 33-mark scripts (from 2005) should win on a lower proportion of occasions against 28-mark scripts (from 2006) than 37-mark scripts (from 2005) should win against 28-mark scripts (from 2006). Also, in any comparison 33-mark scripts (from 2005) should win on a lower proportion of occasions against 34-mark scripts (from 2006) than 37-mark scripts (from 2005) should win against 34-mark scripts (from 2006).

Table 5: The proportion of occasions on which 2005 scripts won against 2006 scripts in the rank ordering conditions

Silent

Proportion of occasions on which 2005 scripts won against 2006 scripts

2005 Mark	33	34	35	36	37
1	30				29
			28*		
0.9					
0.8		29, 33	29	29, 30	30
0.7					
	29, 28*	26*		28*	28*
0.6	33	30, 32	30, 33		32, 33
0.5					
0.4	32		32	33	
0.3					
0.2				32	
0.1					
0					

Thinking Aloud

Proportion of occasions on which 2006 scripts were judged worthy of grade A

2005 Mark	33		34		35		36		37	
	Pk 1	Pk 2	Pk 1	Pk 2	Pk 1	Pk 2	Pk 1	Pk 2	Pk 1	Pk 2
1		30	33	30	31				30	30
				33	33				31	33
									33	
									34	
0.8				32	30	30		30		31
				34						32
										34
0.6			30		34	32		32		
			34			33		34		
						34				
0.4	30	31	31	31	32	31	30	31	32	
	31	32					31			
	31	33					32			
		34					32			
0.2	32		32				34			
0										

* less than 5 participants

How to use Table 5

As an example of how to use Table 5 we can see that the proportion of occasions when 2005 scripts of 33 marks (in the top row) from pack 2 (in the second row from the top) won against 2006 scripts of 30 marks (in the grid) was 1 (in column second from the left); this means the 33-mark 2005 scripts always won.

In Table 5 we would expect to get an approximate increasing monotonic relationship from the bottom left hand corner to the top right hand corner. This is because the proportion of occasions on which 2005 scripts should win against 2006 scripts should increase as the 2005 mark increases, irrespective of the 2006 script mark. For example, in any comparison 33-mark scripts (from 2005) should win against 28-mark scripts (from 2006) on a smaller proportion of occasions than 37-mark scripts (from 2005) should win against 28-mark scripts (from 2006). Also, in any comparison 33-mark scripts (from 2005) should win against 34-mark scripts (from 2006) on a smaller proportion of occasions than 37-mark scripts (from 2005) win against 34-mark scripts (from 2006).

Discussion

Limitations

The experimental conditions reflected current/best practices for awarding, rank ordering and Thurstone pairs procedures within the restrictions of a research study. The first limitation was that the awarding conditions slightly digressed from the awarding practice in two ways:

i. Participants in the experiment did not have any information in addition to scripts that would usually be available in the awarding meeting (apart from the archive scripts, question paper and the mark scheme). This was to avoid influencing the decisions made in the other conditions, which do not include using such information.

ii. Awarders do not always individually make decisions about the quality of candidates' work, although this is not uncommon (Cresswell, 1997). Individual rather than collaborative decisions about individual scripts might increase *if* awarding meetings were undertaken remotely.

Therefore, the awarding conditions in this study might have somewhat limited ecological validity for decisions made silently as well as those made whilst thinking aloud.

A second limitation of our analysis was that the design of the study focused on the main aim of a wider project – to know more about how decisions about grading standards are made from a psychological perspective – and the purpose of the current analysis was of less importance in the study design. There are several aspects to this limitation:

i. The robustness of the statistics is compromised by the small samples of participants and scripts which also affects the generalisability of the study.

ii. The scripts judged silently and whilst thinking aloud were mutually exclusive samples of scripts, consequently, any differences between the modes of grading could be due to the different script samples. The results indicated that the decisions made silently and whilst thinking aloud were broadly similar and therefore this design limitation did not seem to affect the results.

Overall findings

Broadly speaking, verbalising made little difference to the participants' decisions in the various experimental conditions. This is in line with previous research about decisions made silently and whilst thinking aloud in a variety of contexts. Crisp's finding (2008c) that the *marking* decisions made silently are broadly similar to the decisions made whilst thinking

aloud is of particular importance to the assessment community. Our findings are in line with those of Crisp (2008c); these studies reassure us that think aloud concurrent verbal protocols are a robust method for research on decision making in both *marking* and *grading*.

Thus far there is little research literature in the public domain regarding whether the decisions made in Thurstone pairs exercises, rank ordering studies or awarding are the most reliable. The literature also indicates that the visibility of marks can affect decisions in awarding meetings, although there is no similar research for Thurstone pairs and rank ordering. In the present study for all conditions we expected to see an approximate increasing monotonic relationship. For awarding conditions we expect the proportion of occasions on which 2006 scripts are judged worthy of grade A to increase as the 2006 marks increase. For Thurstone pairs and rank ordering conditions we expect the proportion of occasions on which 2005 scripts win against 2006 scripts to increase as the 2005 marks increase. The statistics from the present study suggest that:

- When the scripts are cleaned of marks, the participants make decisions along the expected pattern in Thurstone pairs and rank ordering, but this was not true for the awarding conditions.
- Decisions follow the expected pattern in the Thurstone pairs and awarding conditions with the marks visible. The later can be used to argue that awarding judgements are highly reliable, despite the research literature indicating that the reliability of awarding judgements is less than perfect. On the other hand, it can be argued that given the research literature and the findings of this research, the pattern is almost 'too perfect' for the awarding conditions with marks visible. Perhaps this indicates that the participants relied heavily on the visibility of marks to make their decisions rather than the contents of the scripts or the quality of the candidates' performance.

Laming (2004) theorises that generally, people can make comparisons between two artefacts, but they are not able to maintain a standard in mind and use it to make consistent decisions. This explains why the participants were arguably better at making Thurstone pairs and rank ordering judgements (comparisons between scripts) than making awarding judgements (comparing scripts with internal standards) when the scripts are cleaned of marks. After all, when the marks are visible, decisions can be made based on the marks rather than the examiners' judgements about the gradeworthiness. Indeed, Laming's theory has been used to argue that Thurstone pairs and rank ordering are better methods for maintaining and/or comparing standards than the methods requiring participants to maintain internal standards, for example, at awarding meetings (Bramley, 2005, 2007; Greatorex, 2007; Black and Bramley, 2008).

Implications and recommendations

The present research and other studies (Crisp, 2008c), reassure us that think aloud verbal protocols are a robust research method in the sense that the outcome decisions are unaffected by verbalisation. Therefore, we recommend using concurrent think aloud verbal protocols in future research studies regarding assessment decisions.

This also signifies that our extensive research about the marking and grading judgement processes utilising the method of concurrent think aloud procedure is a trustworthy source of evidence. (See Suto *et al*., 2008) for an overview of the research on the judgement processes in examination marking). The rich qualitative verbal protocol data collected in the main data collection phase of the wider research project are still being analysed and it is hoped that analyses will add to our knowledge about how decisions are made about grading standards.

References

Arlett, S. J. (2003). *A Comparability Study in VCE Health and Social Care, Units 3, 4 and 6: A review of the examination requirements and a report on the cross moderation exercise. A study based on the Summer 2002 examination and organised by AQA on behalf of the Joint Council for General Qualifications*.

Baird, J. (2000). Are examination standards all in the head? Experiments with examiners' judgements of standards in A level examinations. *Research in Education*, **64**, 91–100.

Baird, J. & Scharaschkin, A. (2002). Is the Whole Worth More than the Sum of the Parts? *Studies of Examiners' Grading of Individual Papers and Candidates' Whole A-Level Examination Performances*. *Educational Studies*, **28**, 2, 143–162.

Black, B. & Bramley, T. (2008). Investigating a judgemental rank ordering method for maintaining standards in UK examinations. *Research Papers in Education*, **23**, 3, 357–373.

Bramley, T. (2005). A Rank-Ordering Method for Equating Tests by Expert Judgement. *Journal of Applied Measurement*, **6**, 2, 202–223.

Bramley, T. (2007). Paired Comparison Methods. In: P. Newton, J. Baird, H. Goldstein, H. Patrick & P. Tymms (Eds.), *Techniques for monitoring the comparability of examination standards*. (pp. 246–294) QCA: London.

Bramley, T., Bell, J. F. & Pollitt, A. (1998). Assessing changes in standards over time using Thurstone paired comparisons. *Education Research and Perspectives*, **25**, 2, 1–23.

Cresswell, M. (1997). *Examining Judgements: Theory and Practice of awarding public examination grades*. PhD thesis, University of London Institute of Education: London.

Cresswell, M. (2000). Defining, Setting and Maintaining Standards in Curriculum-Embedded Examinations: Judgemental and Statistical Approaches. In: H. Goldstein & T. Lewis (Eds.), *Assessment: Problems, developments and statistical issues*. (pp. 57 to 84). Chichester: John Wiley and Sons.

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

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

Crisp, V. (2008b). *Judging the grade: An exploration of the judgement processes involved in A level grading decisions*. A paper presented at the British Educational Research Association Conference, September, Heriot-Watt University.

Crisp, V. (2008c). The validity of using verbal protocol analysis to investigate processes involved in examination marking. *Research in Education*, **79**, 1, 1–12.

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

Edwards, E. & Adams, R. (2002). *A Comparability Study in GCE Advanced Level Geography Including the Scottish Advanced Higher Grade Examinations: A review of the examination requirements and a report on the cross moderation exercise. A study based on the Summer 2001 Examination and organised by WJEC on behalf of the Joint Council for General Qualifications*.

Edwards, E. & Adams, R. (2003). *A Comparability Study in GCE Advanced Level Geography Including the Scottish Advanced Higher Grade Examinations: A review of the examination requirements and a report on the cross moderation exercise. A study based on the Summer 2002 Examination and organised by WJEC on behalf of the Joint Council for General Qualifications*.

Ericsson, K. & Simon, H. (1993). *Protocol Analysis: Verbal reports as data*. Cambridge, MA: MIT Press.

Forster, M. & Gray, E. (2000). *Impact of Independent Judges in comparability studies conducted by Awarding Bodies*. A paper presented at the British Educational Research Association Annual Conference, Cardiff University, September.

Good, F. J. & Cresswell M. J. (1988). Grading the GCSE. London: Secondary schools Examination Council. In: M. Cresswell (2000) *Defining, Setting and Maintaining Standards in Curriculum-Embedded Examinations: Judgemental and Statistical Approaches*. In: H. Goldstein & T. Lewis, (Eds.) *Assessment: Problems, developments and statistical issues*. (pp. 57–84). Chichester: John Wiley and Sons.

Greatorex, J. (2003). *What happened to limen referencing? An exploration of how the Awarding of public examinations has been and might be conceptualised*. A paper presented at the British Educational Research Association Conference, 10–13 September 2003 at Heriot-Watt University, Edinburgh.

Greatorex, J. (2007). *Contemporary GCSE and A-level Awarding: A psychological perspective on the decision-making process used to judge the quality of candidates' work*. A paper presented at BERA 2007, University of London.

Greatorex, J., Elliott, G. & Bell, J. F. (2002). *A Comparability Study in GCE AS Chemistry Including parts of the Scottish Higher Grade Examinations: A review of the examination requirements and a report on the cross moderation exercise. A study based on the Summer 2001 examination and organised by the Research and Evaluation Division, University of Cambridge Local Examinations Syndicate for OCR on behalf of the Joint Council for General Qualifications*.

Greatorex, J., Hamnett, L. & Bell, J. F. (2003). *A Comparability Study in GCE Chemistry Including the Scottish Advanced Higher Grade. A study based on the Summer 2002 examination and organised by the Research and Evaluation Division, University of Cambridge Local Examinations Syndicate for OCR on behalf of the Joint Council for General Qualifications*.

Greatorex, J., Novakovic, N. & Suto, I. (2008). *What attracts judges' attention? A comparison of three grading methods*. A paper presented at the IAEA conference, Cambridge.

Green, A. (1998). *Studies in language testing 5: verbal protocol analysis in language testing research*. Cambridge: Cambridge University Press.

Guthrie, K. (2003). *A Comparability Study in GCE Business Studies and VCE Business: A review of the examination requirements and a report on the cross moderation exercise. A study based on the Summer 2002 Examination and organised by the EdExcel on behalf of the Joint Council for General Qualifications*.

Hartog, P. & Rhodes, E. C. (1935). *An Examination of Examinations*. London: Macmillan.

Kimbell, R., Wheeler, A., Miller, S. & Pollitt, A. (2007). *E-scape portfolio assessment phase 2 report*. Department of Design, Goldsmiths, University of London. http://www.goldsmiths.ac.uk/teru/UserFiles/File/e-scape2.pdf

Krahmer, E. & Ummelen, N. (2004). Thinking about thinking aloud: A comparison of two verbal protocols for usability testing. *IEEE Transactions of Professional Communication*, **47**, 2, 105–117.

Laming, D. (2004). *Human judgment: The Eye of the Beholder*. London: Thomson.

Milanovic, M., Saville, N. & Shuhong, S. (1996). A study of decision-making behaviour of composition markers. In: M. Milanovic & N. Saville (Eds.), *Studies in Language Testing 3*. Cambridge: Cambridge University Press.

Murphy, R., Burke, P., Cotton, T., Hancock, J., Partington, J., Robinson, C., Tolley, H., Wilmut, J. & Gower, R. (1995). *The dynamics of GCSE Awarding*. Report of a project conducted for the School Curriculum and Assessment Authority, School of Education, University of Nottingham.

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

Pillner, A. E. G. (1968). Examinations. In: H. J. Butcher (Ed.), *Education Research in Britain*, pp. 167–184. London: University of London Press.

Pinot de Moira, A., Massey, C., Baird, J. & Morrissy, M. (2002). Marking consistency over time, *Research in Education*, **67**, 79–87.

Pollitt, A. & Crisp, V. (2004). *Could comparative judgements of script quality replace traditional marking and improve the validity of examination questions?* A paper presented at the British Educational Research Association, Conference, Manchester.

Pollitt, A. & Elliott, G. (2003a). *Monitoring and Investigating comparability: a proper role for human judgement*. Qualifications and Curriculum Authority Seminar on 'Standards and Comparability', UCLES 4th April 2003.

Pollitt, A. & Elliott, G. (2003b). *Finding a proper role for human judgement in the examination system*. Qualifications and Curriculum Authority Seminar on 'Standards and Comparability', UCLES 4th April 2003.

Qualifications and Curriculum Authority (2008.) *GCSE, GCE, and AEA code of practice 2008*. QCA: London.

Raikes, N. & Massey, A. (2007). Item-level examiner agreement. *Research Matters: A Cambridge Assessment Publication*, **4**, 34–37.

Sanderson, P. J. (2001). *Language and Differentiation in Examining at A level*. PhD thesis, School of Psychology, University of Leeds.

Scharaschkin, A. & Baird, J. (2000). The effects of consistency of performance on A Level examiners' judgements of standards. *British Educational Research Journal*, **26**, 3, 343–357.

Suto, W. M. I., Crisp, V., & Greatorex, J. (2008). Investigating the judgemental marking process. *Research Matters: A Cambridge Assessment Publication*, **5**, 6–8.

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

Suto, W. M. I. & Greatorex, J. (2008b). A quantitative analysis of cognitive strategy usage in the marking of two GCSE examinations. *Assessment in Education: Principles, Policy and Practice*, **15**, 1, 73–89.

Taylor, K. L. & Dionne, J-P. (2000). Accessing problem-solving strategy knowledge: the complementary use of concurrent verbal protocols and retrospective debriefing. *Journal of Educational Psychology*, **92**, 413–425.

Townley, C. (2007). *Australian Education Systems Officials Committee – Secondary Schools Reporting – A study to examine the feasibility of a common scale for reporting all senior secondary subject results*. Victoria Curriculum and Assessment Authority.

Ummelen N., & Neutelings, R. (2000). Measuring reading behavior in policy documents: A comparison of two instruments. *IEEE Trans. Profess. Commun.*, **43**, 3, 292–302. In: E. Krahmer & N. Ummelen (2004), Thinking about thinking aloud: A comparison of two verbal protocols for usability testing. *IEEE Transactions of Professional Communication*, **47**, 2, 105–117.

Van Someren, M., Barnard, Y. & Sandberg, J. (1994). *The think aloud method: a practical guide to modelling cognitive processes*. London: Academic Press.

Vaughan, C. (1992). Holistic assessment: what goes on in the rater's mind? In: L. Hamp-Lyons (Ed.) *Assessing second language writing in academic contexts*. Norwood, NJ: Ablex.

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

Weigle, S. C. (1994). Effects of training on raters of ESL compositions, *Language Testing*, **11**, 197–223.

Willmott, A. S. & Nuttall, D. L. (1975). *The reliability of examinations at 16+*. Schools Council Research Studies. Schools Council Publications. London: MacMillan Education Ltd.

Wilmut, J. (1981). *A Brief Report on two factors which Affect Grade Changes in Mark-Remark and Weighted Exercises*. Associated Examining Board Research report. RAC/184. Guildford: AEB. In: M. Cresswell (2000), Defining, Setting and Maintaining Standards in Curriculum-Embedded Examinations: Judgemental and Statistical Approaches. In: H. Goldstein & T. Lewis (Eds), *Assessment: Problems, developments and statistical issues*. pp. 57–64. Chichester: John Wiley and Sons.

Can emotional and social abilities predict differences in attainment at secondary school?

Carmen L. Vidal Rodeiro, John F. Bell and Joanne L. Emery Research Division

Abstract

Trait Emotional Intelligence (EI) covers a wide range of skills and personality dispositions such as confidence, optimism, adaptability, motivation, peer relations and coping with stress. In recent years the case has been made that emotional and social abilities can be more influential than conventional intelligence for all kinds of personal, career and school success. This study sought to explore the relationship between trait EI and GCSE science performance in a sample of approximately 2000 British students aged 14 to 16. Students were from 31 schools that included both state and independent establishments. The hypothesis was that trait EI would account for better performance at GCSE over and above the level attributable to prior attainment at Key Stage 3.

Trait EI was measured with the Trait Emotional Intelligence Questionnaire: a 153 item, likert-type, self-report instrument that yields a global trait EI score as well as scores for 15 subscales organised into four factors. Participants completed the questionnaire prior to the June 2007 examination session and their responses were matched to their Key Stage 3 and GCSE results. Attainment in different GCSE science subjects was modelled through separate regression analyses.

Results showed that some aspects of trait EI significantly predicted attainment in GCSE sciences over and above the contribution made by Key Stage 3 attainment. The majority of the questionnaire subscales significantly predicted attainment in the Applied Science Double Award after controlling for Key Stage 3 scores. Self-motivation and low impulsivity were significant predictors of attainment in all of the science subjects here after controlling for Key Stage 3 scores. Global trait EI scores significantly predicted progress from Key Stage 3 in the Applied Science Double Award and in Biology and Chemistry but not in Physics.

Introduction

One piece of evidence that is used by awarding bodies when setting pass marks for school examinations in England is the prior attainment of the candidates. It is not unreasonable to expect that examination results will improve if the prior attainment of the candidates improves from that of the previous year. However, prior attainment is not the only determinant of examination performance. This can be illustrated by considering what happened when vocational GCSEs (GCSE(v)) were introduced in England. These examinations were introduced to give a more practical alternative to the academic GCSE examinations. It was hoped that this would improve the motivation of these students. When the first results were released concern was expressed that the grades tended to be lower than expected given candidates' attainment at age 11. A thorough analysis revealed that the candidates also made less progress than expected from National Tests at age 14. However, there was no evidence that the pupils' results in GCSE(v)s tended to be any lower than in their other GCSE subjects (that is, they also made less progress than expected in their non-vocational GCSEs). It was thought that a possible reason for this was that the GCSE(v) candidates tended to be less motivated (Vidal Rodeiro and Bell, 2007).

The objective of this study was therefore to investigate whether relationships exist between the affective domain and progress in school. After reviewing the affective literature it was decided that an investigation into emotional intelligence might provide an insight into the reasons for differential progress in schools. This involves attributes such as motivation, stress management and self-control: factors which could conceivably influence school performance in addition to ability. This study was designed to investigate the following research questions:

1. Do the entries of different OCR science specifications (i.e. the sets of candidates taking the examinations) vary in their emotional intelligence?
2. Can this variation be accounted for by variation in prior attainment?
3. Is progress on the different science specifications associated with candidates' levels of emotional intelligence?

If the answers to all of these questions are 'yes' then it would suggest that care needs to be taken when using prior attainment to predict performance in the processes of setting and maintaining examination standards. It would also suggest that, if attempts to develop the emotional intelligence of schoolchildren prove to be successful, then these would be worthwhile provided that the relationship between EI and examination success is a causal one. This will be discussed later.

National Curriculum subjects such as PSE/PSHE and Citizenship target pupils' social, emotional and behavioural skills. Many primary and secondary schools are currently using new curriculum materials for actively developing their pupils' social, emotional and behavioural skills (DfES, 2005, 2007). An example of this is the 'Social and Emotional Aspects of Learning' program (SEAL), which is a comprehensive approach to promoting the social and emotional skills that underpin effective learning, positive behaviour, regular attendance, staff effectiveness and the emotional health and well-being of all who learn and work in schools. It is argued that the social and emotional aspects of learning, such as self-awareness, managing feelings, motivation, empathy, and social skills, are key areas that can and need to be developed in children so that they can learn effectively. Research has suggested that motivation, along with abilities and other personality traits, is important in predicting academic school performance (e.g. Abouserie, 1995; Gumora and Arsenio, 2002; Lam and Kirby 2002; Humphrey *et al*., 2007).

This study uses a questionnaire that measures trait emotional intelligence. Goleman (1996) popularised the term 'emotional intelligence' and argued that emotional and social abilities can be more influential than conventional intelligence for all kinds of personal, career and school success. The definitions of emotional intelligence are varied and researchers are constantly amending definitions of the construct

(Dulewicz and Higgs, 2000). In this research the Petrides and Furnham (2000) model is used. This proposes a conceptual distinction between the ability-based model and the trait-based model of emotional intelligence. Their trait emotional intelligence (or 'trait emotional self-efficacy') is defined as:

> *a constellation of behavioral dispositions and self-perceptions concerning one's ability to recognize, process, and utilize emotion-laden information.* (Petrides and Furnham, 2000)

Trait emotional intelligence (trait EI) is regarded by these authors as a dimension of personality rather than a form of intelligence due to its relationship with certain personality traits and its lack of a relationship with non-verbal reasoning ability (Petrides and Furnham, 2000; Petrides *et al*., 2004).

This study explored the relationships between trait EI and academic performance in a sample of British students. It investigated whether trait EI accounts for better performance in examinations at age 16 over and above the level predicted by prior attainment at age 14.

Method

Trait EI was measured with the Trait Emotional Intelligence Questionnaire (TEIQue v. 1.50): a likert-type, self-report instrument devised and developed by Petrides (2001) and Petrides and Furnham (2003). As a self-report instrument, the TEIQue measures people's perceptions of their own abilities.

The version of the questionnaire used in this research has 153 items and yields a global score as well as scores for each of 15 subscales organised into four factors. Table 1 lists the 15 trait EI subscales along with a brief description of each of them.

Table 1: Emotional intelligence subscales

Subscale	*High scorers perceive themselves as…*	
Self-esteem	…successful and self-confident.	1
Emotion expression	…capable of communicating their feelings to others.	2
Self-motivation	…driven and unlikely to give up in the face of adversity.	3
Emotion regulation	…capable of controlling their emotions.	4
Happiness	…cheerful and satisfied with their lives.	5
Empathy	…capable of taking someone else's perspective.	6
Social awareness	…accomplished networkers with excellent social skills.	7
Impulsivity (low)	…reflective and less likely to give in to their urges.	8
Emotion perception	…clear about their own and other people's feelings.	9
Stress management	…capable of withstanding pressure and regulating stress.	10
Emotion management	…capable of influencing other people's feelings.	11
Optimism	…confident and likely to "look on the bright side" of life.	12
Relationships	…capable of having fulfilling personal relationships.	13
Adaptability	…flexible and willing to adapt to new conditions.	14
Assertiveness	…forthright, frank, and willing to stand up for their rights.	15

The TEIQue also provides scores on four factors:

- **Wellbeing:** a combined score of optimism, happiness and self-esteem.
- **Self-control:** a combined score of emotion regulation, impulsiveness and stress management.
- **Emotionality:** a combined score of empathy, emotion perception, emotion expression and relationships.
- **Sociability:** a combined score of emotion management, assertiveness and social awareness.

All TEIQue scores (subscales, factors and global) vary between 1 and 7 and higher scores indicate higher levels of trait emotional intelligence. Descriptive statistics providing the mean values and the standard deviations of each of the TEIQue subscales in this sample are given in Table 2.

Table 2: Means and standard deviations of the TEIQue subscales

Variable	*Mean*	*Standard Deviation*	*Minimum*	*Maximum*
Self-esteem	4.47	1.04	1.00	7.00
Emotion expression	4.45	1.04	1.00	7.00
Self-motivation	4.31	0.84	1.20	6.90
Emotion regulation	3.93	0.85	1.08	7.00
Happiness	5.22	1.20	1.00	7.00
Empathy	4.63	0.85	1.33	7.00
Social awareness	4.65	0.83	1.00	7.00
Impulsivity (low)	3.94	0.94	1.00	7.00
Emotion perception	4.57	0.79	1.40	7.00
Stress management	4.16	0.96	1.10	7.00
Emotion management	4.66	0.84	1.00	7.00
Optimism	4.94	1.03	1.00	7.00
Relationships	5.17	0.84	1.44	7.00
Adaptability	4.17	0.75	1.56	6.78
Assertiveness	4.61	0.93	1.00	7.00
Wellbeing	4.88	0.96	1.46	7.00
Self-control	4.01	0.75	1.24	6.56
Emotionality	4.71	0.66	1.66	6.75
Sociability	4.64	0.73	1.04	6.85
Trait EI	**4.53**	**0.57**	**2.29**	**6.59**

Two hundred and fourteen schools were invited to take part in the research. The questionnaire was administered in the period immediately before the GCSE examinations were to be taken. Unfortunately, this might have been the reason why the response rate was relatively low (many schools turned down the opportunity to take part although the vast majority of eligible pupils within the participating schools returned a questionnaire). Although a small proportion of questionnaires was incomplete, the final sample comprised 1977 students in 31 schools who were taking OCR[1] GCSE science exams in June 2007. All participants were in Year 10 or Year 11 of school. It should be noted that the study was designed to compare the different science specifications and was restricted to OCR science examinations. This means that the resulting sample was not intended to be representative of the whole population. In particular, the proportion of candidates entered for separate sciences and attending independent schools was higher than in the whole population.

The examination most commonly taken at the end of Key Stage 4 is the General Certificate of Secondary Education (GCSE). There are eight grades: A*, A, B, C, D, E, F and G. Students who fail to reach grade G are recorded as U (unclassified). Students were invited to participate in this study if they were entered for an examination in at least one of the following OCR science subjects: Applied Science Double Award, Biology,

1 Oxford, Cambridge and RSA Examinations

Physics, Chemistry, Science: Double Award, Science: Twenty First Century Science Suite and Science: Gateway Science Suite. The last two specifications are modular and the candidates taking these in this study were all in Year 10. Unfortunately, the response rate for Science Double Award was too low to allow meaningful analysis. This article therefore concentrates on the remaining four specifications: Applied Science Double Award (vocational) and the three separate sciences.

The separate sciences (Biology, Chemistry and Physics) were usually taken by the same candidates: only a small number here did not take all three subjects. Nobody taking the vocational science subject took any of the separate science subjects. Many of the pupils in the sample were tested at age 14 (Key Stage 3) and were awarded attainment levels ranging from 1 to 8. These tests cover English, Mathematics and Science. The total of the levels is used as the prior attainment variable in this study. Around 30% of the sample did not take Key Stage 3 tests (students at independent schools are not required to). Of the separate sciences candidates with Key Stage 3 scores, around a third were female and around two thirds were male (for all three subjects).

Results

The aim of the survey was to investigate the relationships between EI and particular OCR specifications. The initial study design meant that more centres were asked to take part from some specification types than others. For example, the three separate sciences are much more likely to be taken in independent and grammar schools. The lower than hoped for participation rate by schools led to a distribution of school types that severely restricted the analyses that could be done at the school level due to the small number of schools in each cell (see Table 3). In addition, it became clear in exploratory data analysis that the single girls-only grammar school had particularly low values on some of the EI factors. This school had an OFSTED inspection two months after the questionnaire was completed. This report noted that the school was recovering from difficulties which were not specified. However, there was a quote from a pupil attending the school that the atmosphere was improving day by day.

Table 3: The distribution of school types taking part in the study

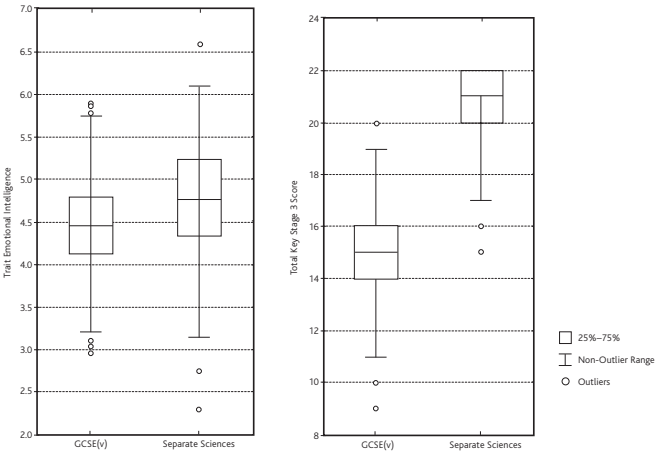
School Type	*Boarding*	*Boys*	*Girls*	*Mixed*	*Grand Total*
Comprehensive	No		4	14	18
Grammar	No	2	1		3
Independent	No		2	1	3
Independent	Yes	1		2	3
Independent Total		1	2	3	6
Secondary Modern	No		2	2	4
Grand Total		**3**	**9**	**19**	**31**

When scores on the trait EI subscales were compared for the Applied Science GCSE(v) and the separate sciences entries (Table 4) it was found that, for all subscales except emotion expression and optimism, the mean scores for the Applied Science GCSE(v) entry were significantly lower than those for the separate sciences. In addition, the performance of the separate sciences entry at Key Stage 3 was considerably higher. The entry of the vocational GCSE tends to be composed of much lower performers at Key Stage 3 than the entry for the separate sciences (as illustrated in the box plots in Figure 1). This has the implication that the relationships between attainment and the trait EI scales for the vocational science and for the separate sciences will apply to different parts of the attainment range. If there is any non-linearity in the relationships between attainment and the trait EI scales then different results may be expected between the vocational science subject and the separate science subjects.

Table 4: Comparison of mean EI scores and Key Stage 3 performance for Applied Science GCSE(v) and the separate sciences entries

Subscale	*Mean Applied Science*	*Mean Separate Sciences*	*t-value*	*df*	*p*
Self-esteem	4.50	4.74	-3.27	723	0.00
Emotion expression	4.46	4.48	-0.32	723	0.75
Self-motivation	4.27	4.50	-3.80	723	0.00
Emotion regulation	3.94	4.33	-6.52	723	0.00
Happiness	5.07	5.44	-4.21	723	0.00
Empathy	4.56	4.89	-5.56	723	0.00
Social awareness	4.55	4.89	-5.52	723	0.00
Impulsivity (low)	3.94	4.21	-3.97	723	0.00
Emotion perception	4.55	4.73	-3.15	723	0.00
Stress management	4.08	4.53	-6.39	723	0.00
Emotion management	4.45	4.95	-8.16	723	0.00
Optimism	4.93	5.05	-1.59	723	0.11
Relationships	5.13	5.30	-2.88	723	0.00
Adaptability	4.13	4.35	-3.95	723	0.00
Assertiveness	4.51	4.89	-5.72	723	0.00
Wellbeing	4.83	5.08	-3.45	723	0.00
Self-control	3.99	4.36	-6.84	723	0.00
Emotionality	4.67	4.85	-3.66	723	0.00
Sociability	4.51	4.91	-7.64	723	0.00
Trait EI	4.47	4.75	-6.67	723	0.00
Total Key Stage score	**14.92**	**20.89**	**-40.06**	**574**	**0.00**

Figure 1: Box plots of measures by science entries



(a) Trait Emotional Intelligence (b) Total Key Stage 3 Score

GCSE(v) Double Award in Applied Science

283 students in the survey sat a GCSE(v) Double Award in Applied Science and had a Key Stage 3 score. The grades obtained ranged from AA to GG with CC being the modal grade. This set of students was quite different to the set taking the separate sciences. For example, only around 3% of these students obtained at least a grade AA (compared

with 75% of students in the sample obtaining at least a grade A in Biology). This is to be expected given the difference in prior attainment at Key Stage 3.

In a proportional odds model the probability of obtaining at least a grade k is given by the following equation:

$$\ln \frac{\pi_k}{1-\pi_k} = \alpha_k + \beta x$$

where α_k is a constant for grade k and β is the slope for the Key Stage 3 score, x.

Proportional odds models were used as there was no significant evidence of non-proportionality in any of the analyses (that is, different slopes for each grade) but, given the distribution of grades and the sample sizes, any difference would have to be large to be detected.

In Table 5 the parameters for the independent variables are given for GCSE(v) Applied Science. Each EI subscale was modelled separately. The estimates represent the log of the odds ratio of attaining at least a particular GCSE grade. All significant effects are highlighted in bold type (an estimate is statistically significant if it equals twice or more the value of the standard error). A positive significant gender effect indicates that, for given values of the EI subscale in the model and a given Key Stage 3 score, the probability of obtaining at least a given grade is higher for females than for males. This was the case for the self-motivation, emotion regulation and stress management subscales, and for the self-control factor.

A positive significant EI subscale effect indicates that, for a given Key Stage 3 score, the probability of obtaining at least a given grade significantly increases with increasing scores on that subscale. It can be seen in Table 5 that most of the EI subscales had a positive relationship with the probability of obtaining at least a given grade in this subject when Key Stage 3 performance was controlled for. The only exceptions were the emotion expression, emotion management and assertiveness subscales and the sociability factor.

Figure 2 illustrates that a male candidate with a total Key Stage 3 score of 16 and an overall trait EI score of 3 would have a predicted probability of obtaining at least a grade CC of 0.42. If that same candidate's trait EI score was 6 then their predicted probability would be 0.92. A more modest difference in trait EI from 3 to 4 would increase the predicted probability of obtaining a grade CC from 0.42 to 0.63. If this is a causal relationship, where changes in an individual's trait EI changes their probability of success in examinations (given that one of the subscales is self-motivation this is plausible), then the performance of school children could be improved substantially by devising strategies for even modest improvements in their emotional intelligence.

GCSE Biology

244 students in the sample took the Biology GCSE and had a total Key Stage 3 score. The grades obtained were A* to D with A being the modal grade (such a small grade range is to be expected since the separate sciences are usually taken by relatively high achievers). In Table 6 the parameters for the independent variables are given for GCSE Biology. For most of the subscales the gender effect was positive and significant. The exceptions were the emotion expression, empathy, emotion management and relationships subscales. The self esteem, self motivation, happiness, empathy, low impulsivity, relationships and adaptability subscales, the wellbeing and self-control factors and the global score were all significant

Table 5: Proportional odds regression parameters for gender, total Key Stage 3 score and the emotional intelligence subscales for GCSE(v) Applied Science

Subscale	*Gender (=F)*		*EI subscale*		*Total KS3 score*	
	Estimate	*Std Err*	*Estimate*	*Std Err*	*Estimate*	*Std Err*
Self-esteem	0.22	0.12	**0.29**	**0.12**	**0.49**	**0.06**
Emotion expression	0.18	0.12	0.06	0.12	**0.47**	**0.06**
Self-motivation	**0.24**	**0.12**	**0.59**	**0.16**	**0.48**	**0.06**
Emotion regulation	**0.27**	**0.12**	**0.47**	**0.15**	**0.47**	**0.06**
Happiness	0.19	0.12	**0.24**	**0.09**	**0.46**	**0.06**
Empathy	0.15	0.12	**0.40**	**0.15**	**0.46**	**0.06**
Social awareness	0.20	0.12	**0.30**	**0.15**	**0.48**	**0.06**
Impulsivity (low)	0.20	0.12	**0.69**	**0.14**	**0.51**	**0.06**
Emotion perception	0.17	0.12	**0.53**	**0.16**	**0.48**	**0.06**
Stress management	**0.28**	0.12	**0.43**	**0.12**	**0.47**	**0.06**
Emotion management	0.19	**0.12**	-0.06	0.14	**0.47**	**0.06**
Optimism	0.22	0.12	**0.30**	**0.12**	**0.48**	**0.06**
Relationships	0.12	0.12	**0.50**	**0.15**	**0.49**	**0.06**
Adaptability	0.22	0.12	**0.29**	**0.12**	**0.47**	**0.06**
Assertiveness	0.19	0.12	0.15	0.14	**0.47**	**0.06**
Wellbeing	0.22	0.12	**0.35**	**0.13**	**0.49**	**0.06**
Self-control	**0.30**	**0.12**	**0.81**	**0.17**	**0.48**	**0.06**
Emotionality	0.14	0.12	**0.65**	**0.20**	**0.47**	**0.06**
Sociability	0.19	0.12	0.18	0.18	**0.47**	**0.06**
Trait EI	0.23	**0.12**	**0.93**	**0.23**	**0.48**	**0.06**

(Full details of all the models can be obtained from the authors)

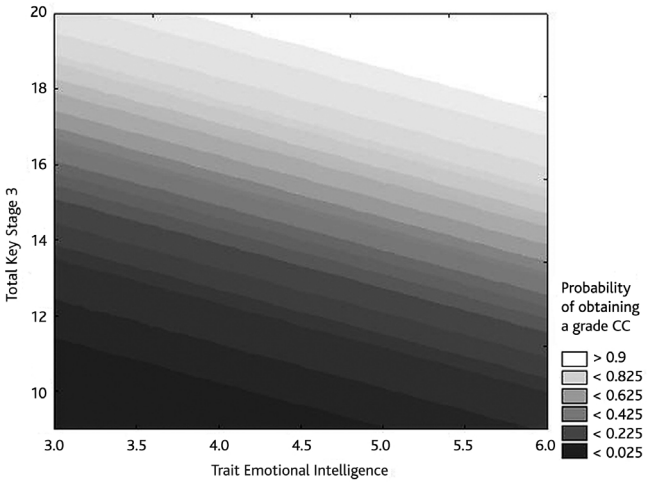


Figure 2: Predicted probability of a male candidate obtaining at least a grade CC in GCSE(v) Double Award in Applied Science

predictors of attainment in Biology when controlling for Key Stage 3 attainment. Prior attainment was a much more powerful predictor than was the case for Applied Science but it should be noted that the two sets of data differ considerably in their prior attainment scores and that the relationships therefore refer to different parts of the attainment range.

GCSE Chemistry

For GCSE Chemistry there were 241 candidates with valid Key Stage 3 scores. Again the grades ranged from A* to D. However, in this case the modal grade was A*. Table 7 gives the parameters for the independent variables for GCSE Chemistry. For the self-esteem and adaptability subscales, and for the wellbeing factor, there was a gender effect in favour of females. The following subscales and factors were related to improved performance in Chemistry when controlling for Key Stage 3 attainment: self-esteem, self motivation, happiness, low impulsivity, optimism, adaptability, wellbeing, self-control and the global score. Key Stage 3 performance was a strong predictor of performance in this subject.

Table 6: Proportional odds regression parameters for gender, total Key Stage 3 score and the emotional intelligence subscales for GCSE Biology

Subscale	*Gender (=F)*		*EI subscale*		*Total KS3 score*	
	Estimate	*Std Err*	*Estimate*	*Std Err*	*Estimate*	*Std Err*
Self-esteem	**0.50**	**0.16**	**0.39**	**0.13**	**0.81**	**0.11**
Emotion expression	0.25	0.14	-0.10	0.11	**0.83**	**0.11**
Self-motivation	**0.39**	**0.14**	**0.61**	**0.14**	**0.80**	**0.11**
Emotion regulation	**0.35**	**0.15**	0.25	0.16	**0.79**	**0.11**
Happiness	**0.39**	**0.15**	**0.34**	**0.10**	**0.84**	**0.11**
Empathy	0.22	0.14	**0.37**	**0.15**	**0.81**	**0.11**
Social awareness	**0.30**	**0.14**	0.17	0.13	**0.82**	**0.11**
Impulsivity (low)	**0.30**	**0.14**	**0.64**	**0.13**	**0.77**	**0.10**
Emotion perception	**0.30**	**0.14**	0.13	0.14	**0.82**	**0.11**
Stress management	**0.28**	**0.14**	0.05	0.13	**0.81**	**0.11**
Emotion management	0.26	0.14	-0.09	0.14	**0.82**	**0.11**
Optimism	**0.34**	**0.15**	0.19	0.11	**0.82**	**0.11**
Relationships	0.23	0.14	**0.54**	**0.15**	**0.81**	**0.11**
Adaptability	**0.50**	**0.16**	**0.39**	**0.13**	**0.81**	**0.11**
Assertiveness	**0.28**	**0.14**	0.06	0.12	**0.82**	**0.11**
Wellbeing	**0.43**	**0.15**	**0.38**	**0.13**	**0.83**	**0.11**
Self-control	**0.38**	**0.15**	**0.49**	**0.17**	**0.77**	**0.10**
Emotionality	**0.28**	**0.14**	0.29	0.17	**0.81**	**0.11**
Sociability	**0.28**	**0.14**	0.07	0.15	**0.82**	**0.11**
Trait EI	**0.39**	**0.15**	**0.59**	**0.20**	**0.80**	**0.11**

Table 7: Proportional odds regression parameters for gender, total Key Stage 3 score and the emotional intelligence subscales for GCSE Chemistry

Subscale	*Gender (=F)*		*EI subscale*		*Total KS3 score*	
	Estimate	*Std Err*	*Estimate*	*Std Err*	*Estimate*	*Std Err*
Self-esteem	**0.40**	**0.16**	**0.47**	**0.13**	**0.99**	**0.12**
Emotion expression	0.08	0.14	-0.13	0.12	**1.00**	**0.12**
Self-motivation	0.22	0.14	**0.53**	**0.14**	**0.97**	**0.11**
Emotion regulation	0.24	0.15	0.33	0.17	**0.97**	**0.12**
Happiness	0.22	0.15	**0.27**	**0.10**	**1.00**	**0.12**
Empathy	0.11	0.14	0.08	0.15	**0.99**	**0.12**
Social awareness	0.15	0.14	0.13	0.13	**0.99**	**0.12**
Impulsivity (low)	0.13	0.14	**0.58**	**0.14**	**0.95**	**0.11**
Emotion perception	0.13	0.14	0.04	0.14	**0.99**	**0.12**
Stress management	0.15	0.15	0.09	0.13	**0.98**	**0.12**
Emotion management	0.11	0.14	-0.04	0.15	**0.99**	**0.12**
Optimism	0.21	0.15	**0.22**	**0.11**	**1.00**	**0.12**
Relationships	0.11	0.14	0.21	0.15	**0.98**	**0.12**
Adaptability	**0.40**	**0.16**	**0.47**	**0.13**	**0.99**	**0.12**
Assertiveness	0.15	0.14	0.19	0.12	**0.99**	**0.12**
Wellbeing	**0.30**	**0.15**	**0.38**	**0.13**	**1.00**	**0.12**
Self-control	0.24	0.15	**0.52**	**0.18**	**0.95**	**0.11**
Emotionality	0.12	0.14	0.04	0.18	**0.99**	**0.12**
Sociability	0.15	0.14	0.15	0.15	**0.99**	**0.12**
Trait EI	0.25	0.15	**0.57**	**0.21**	**0.98**	**0.12**

GCSE Physics

GCSE Physics had the fewest candidates in the sample with a valid Key Stage 3 score. Data from 225 candidates were analysed. The grades ranged from A* to E with A* being the modal grade. Table 8 gives the parameters for the independent variables for GCSE Physics. For all subscales here the effect of female gender was negative (although not significantly so for self-esteem, emotion regulation and adaptability). Only two of the EI subscales had a significant relationship with GCSE performance after controlling for Key Stage 3 attainment. These were self-motivation and low impulsivity. Assuming causality, for a candidate with a Key Stage 3 score of 21 an increase on the self-motivation scale from 4 to 5 would increase their predicted probability of getting an A* grade from 0.5 to 0.58. Of all the science subjects here, Key Stage 3 scores had the strongest influence on Physics performance.

Table 8: Proportional odds regression parameters for gender, total Key Stage 3 score and the emotional intelligence subscales for GCSE Physics

Subscale	*Gender (=F)*		*EI subscale*		*Total KS3 score*	
	Estimate	*Std Err*	*Estimate*	*Std Err*	*Estimate*	*Std Err*
Self-esteem	-0.34	0.18	0.17	0.15	**1.02**	**0.12**
Emotion expression	**-0.48**	**0.16**	-0.11	0.14	**1.03**	**0.12**
Self-motivation	**-0.38**	**0.16**	**0.32**	**0.15**	**1.02**	**0.12**
Emotion regulation	-0.32	0.17	0.33	0.19	**1.01**	**0.12**
Happiness	**-0.38**	**0.16**	0.15	0.12	**1.03**	**0.12**
Empathy	**-0.44**	**0.16**	-0.11	0.19	**1.03**	**0.12**
Social awareness	**-0.46**	**0.16**	-0.07	0.17	**1.03**	**0.12**
Impulsivity (low)	**-0.42**	**0.16**	**0.48**	**0.17**	**1.01**	**0.12**
Emotion perception	**-0.46**	**0.16**	-0.05	0.18	**1.03**	**0.12**
Stress management	**-0.45**	**0.16**	-0.01	0.16	**1.03**	**0.13**
Emotion management	**-0.50**	**0.16**	-0.23	0.18	**1.01**	**0.12**
Optimism	**-0.43**	**0.17**	0.03	0.13	**1.03**	**0.12**
Relationships	**-0.46**	**0.16**	0.27	0.18	**1.02**	**0.12**
Adaptability	-0.34	0.18	0.17	0.15	**1.02**	**0.12**
Assertiveness	**-0.45**	**0.16**	-0.02	0.15	**1.03**	**0.12**
Wellbeing	**-0.38**	**0.17**	0.15	0.15	**1.03**	**0.12**
Self-control	**-0.35**	**0.16**	0.38	0.21	**1.00**	**0.12**
Emotionality	**-0.45**	**0.16**	-0.02	0.22	**1.03**	**0.12**
Sociabiliy	**-0.48**	**0.16**	-0.13	0.19	**1.02**	**0.12**
Trait EI	**-0.38**	**0.17**	0.24	0.25	**1.02**	**0.12**

Discussion

Emotional intelligence currently attracts a great deal of interest, both in academia and with the general public. In education it has been claimed that people with high scores on a trait EI measure perform better at school (e.g. Lam and Kirby, 2002; Petrides *et al*., 2004; Zins *et al*., 2004). The present study provides support for the role of trait EI in students' performance and progress at secondary school.

Factors such as ability are not the only predictors of educational attainment. According to this study, and also according to previous research (Cassidy and Lynn, 1991; Vidal Rodeiro and Bell, 2007), it is the combination of ability, individual characteristics, home background, type of school attended and social, behavioural and emotional aspects that is important.

The results show that some aspects of trait emotional intelligence were significantly related to attainment in GCSE sciences over and above the contribution made by prior ability (Key Stage 3 scores). Self-motivation and low impulsivity were significant positive predictors of progress from Key Stage 3 in all four science subjects here. On the other hand, the emotion expression, emotion management and assertiveness subscales, and the sociability factor, were not significant predictors of progress in any of them. These findings corroborate those of Petrides *et al*. (2004), who found that EI moderated the relationship between cognitive ability and performance. Similarly, Gumora and Arsenio (2002) found that some aspects of EI contributed to performance at school over and above the contribution made by cognition-related abilities.

In this research the relationships between trait EI and performance in four different science subjects at GCSE were studied. Some GCSE subjects appear to require more consideration of affect-related issues (e.g. English Literature, Art, Drama, etc.) and therefore trait EI may be found to be a

better predictor of performance in some subjects than in others. Petrides *et al*. (2004) found a differential influence of trait EI on Mathematics, English and Science attainment. A future intention with this research is to match all the GCSE results of the participants to their trait EI scores in order to investigate the relationships between trait EI and performance in a wide range of GCSE subjects.

The results of this study show that trait EI was differentially implicated in academic progress across the various GCSE science subjects considered and influenced progress from Key Stage 3 in some more than in others. Trait EI scores had the greatest effect on progress in the Applied Science Double Award and the least effect on progress in Physics. The predictiveness of Key Stage 3 attainment was lowest for the Applied Science Double Award and highest for Physics. There are large differences in the prior attainment of the entries for these examinations and this suggests a possibly non-linear relationship between trait EI and progress over the range of prior attainment. That is, trait EI may have a larger effect where prior attainment is lower and a smaller effect where prior attainment is higher.

Schools and students were self-selected for this study and this might be a limitation since it is possible that the more able and/or confident students would have been more likely to complete the questionnaire. Also, schools that were more involved in the promotion of EI ideas might have been more likely to take part. Finally, the present study was limited by being restricted to students taking science subjects. Further research on the long-term stability of trait EI may also be of interest.

Earlier in the article it was suggested that if there is a causal relationship between emotional intelligence and examination performance then the results of this study suggest that substantial improvements in attainment are possible if emotional intelligence can be raised. Emotional intelligence is a relatively new way of considering the affective domain. The latter term was developed by Bloom in his *Taxonomy of Educational Objectives* (Bloom *et al*., 1956). The top level of this classification had three categories: cognitive, affective and psychomotor. Loosely, the first is the thinking skills used in learning and the third describes the ability to physically manipulate a tool or instrument, for example, you cannot teach a child to write if they have not develped the skills to control a pencil. It is the second domain in which this research is focussed. This domain includes the manner in which we deal with things emotionally, such as feelings, values, appreciation, enthusiasms, motivations and attitudes.

The importance of the affective domain in education has long been recognised. For example, Thomas Arnold, the famous headmaster of Rugby School, believed that, while learning was important, the great aim of education was the formation of character. His ideal was to train boys to become not merely scholars but Christian gentlemen. After allowing for the mores and language of the era it is clear that features of emotional intelligence, such as adaptability, emotion management, low impulsiveness, self-motivation and social awareness, were meant to be developed. Today Rugby School's website states: 'Many fundamental qualities are not examinable: curiosity, shrewdness, initiative, an awareness of beauty, a sense of humour, a sense of responsibility and a gift for friendship. These qualities need to be developed in an institution that regards itself as educational….'

The components measured in trait emotional intelligence have existed previously as part of other questionnaires and similar factors have long been measured in the affective domain, although there may be differences in the precise wording. For example, emotional regulation is a very similar concept to emotional resilience and is not unrelated to the nineteenth century concept of 'stiff upper lip'. There is thus a considerable body of research evidence relevant to establishing a causal relationship between emotional intelligence and educational attainment. For example, by gathering evidence from sixty one research experts, ninety one formal review papers and one hundred and seventy nine handbook chapters, Wang *et al*. (1993) found that the 'affective-motivational attitudinal disposition of students' was more important than peer group, school culture and the quantity and quality of classroom instruction in influencing learning outcomes. Focussing solely on curriculum and teaching initiatives might therefore not be the most effective way of improving examination performance. It is also worth noting that it is not unreasonable to expect the quality of instruction to be positively related to the levels of emotional intelligence of the students.

More recent research findings have supported the argument that features of the affective domain have a particular and separate impact on achievement. Some of the most useful research in this area is the review of positive youth development programs in the United States by Richard Catalano and his colleagues. They obtained a consensus that positive youth development programs sought to achieve one or more of the following objectives: promotes bonding, fosters resilience, promotes social competence, promotes emotional competence, promotes behavioural competence, fosters self determination, fosters spirituality, fosters self-efficacy, fosters clear and positive identity, fosters belief in the future, provides recognition of positive behaviours, fosters opportunities for pro-social involvement and fosters pro-social norms. Again the words may be different but many of the ideas are the same as those used in emotional intelligence.

Using very rigorous criteria for identifying effective programs, Catalano *et al*. (2004) identified thirty studies that could be used to draw sound conclusions about the effects on youth's behavioural and educational outcomes. Twenty five of these programs were successful. Nineteen of the programs showed significant improvements in a range of factors including interpersonal skills, quality of relationships, self control and academic achievement. They concluded that it was schemes involving methods that, in effect, improved emotional intelligence that produced these benefits. They also concluded that a structured programme is more likely to be a success and that it needs to be clear and well planned. They noted that structured programs that included opportunities to practice skills and gave feedback and positive reinforcement were more likely to be successful.

Another example of this type of work that has been evaluated is the Australian 'You can do it!' Programme (Bernard, 2006). This research found that, in another variant terminology, academic confidence, work persistence, work organisation, getting along and emotional resilience can be taught. Not only can these be taught but, following the training, academic performance is increased. The aim of this program is to create beneficial habits of mind, defined as an automatic tendency of a person to think in a certain way.

In conclusion, the research supports the premise that emotional intelligence has a very important effect on learning and that it is possible to improve it with training programs. In particular, it may be more effective than concentrating solely on teaching and curriculum initiatives.

References

Abouserie, R. (1995). Self-esteem and achievement motivation as determinants of students' approaches to studying. *Studies in Higher Education*, **20**, 1, 19–26.

Bernard, M. (2006). It's time we teach social-emotional competence as well as we teach academic competence. *Reading and Writing Quarterly*, **22**, 2, 103–119.

Bloom, B.S. (Ed.), Engelhart, M.D., Furst, E.J., Hill, W.H., & Krathwohl, D.R. (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook 1: Cognitive domain*. New York: David McKay.

Catalano, R. F., Berglund, M. L., Ryan, J. A. M., Lonczak, H. S. & Hawkins, J. D. (2004). Positive Youth Development in the United States: Research Findings on Evaluations of Positive Youth Development Programs. *American Academy of Political and Social Science*, **591**, 98–124.

Cassidy, T. & Lynn, R. (1991). Achievement motivation, educational attainment, cycles of disadvantage and social competence: some longitudinal data. *British Journal of Educational Psychology*, **61**, 1–12.

Department for Education and Skills (2005). *Social and emotional aspects of learning: guidance*. DfES report 1378–2005. London: DfES.

Department for Education and Skills (2007). *Social and emotional aspects of learning for secondary schools (SEAL). Guidance Book*. London, DfES.

Dulewicz, V. & Higgs, M. (2000). Emotional intelligence – A review and evaluation study. *Journal of Managerial Psychology*, **15**, 4, 341–372.

Goleman, D. (1996). *Emotional Intelligence: why it can matter more than IQ*. New York: Bantam Books.

Gumora, G. & Arsenio, W.F. (2002). Emotionality, emotion regulation and school performance in middle school children. *Journal of School Psychology*, **40**, 5, 395–413.

Humphrey, N., Curran, A., Morris, E., Farrell, P. & Woods, K. (2007). Emotional Intelligence and Education: A critical review. *Educational Psychology*, **27**, 2, 235–254.

Lam, L.T. & Kirby, S.L. (2002). Is Emotional Intelligence an Advantage? An exploration of the impact of Emotional Intelligence on individual performance. *Journal of Social Psychology*, **142**, 1, 133–143.

Petrides K. V. (2001). *A psychometric investigation into the construct of emotional intelligence*. Doctoral dissertation. University College London, London, England.

Petrides, K.V., Frederickson, N. & Furnham, A. (2004). The role of trait emotional intelligence in academic performance and deviant behaviour at school. *Personality and Individual Differences*, **36**, 277–293.

Petrides, K. V. & Furnham, A. (2000). On the dimensional structure of emotional intelligence. *Personality and Individual Differences*, **29**, 313–320.

Petrides, K. V. & Furnham, A. (2003). Trait emotional intelligence: Behavioural validation in two studies of emotion recognition and reactivity to mood induction. *European Journal of Personality*, **17**, 39–57.

Petrides, K.V., Furnham, A. & Martin, N.G. (2004). Estimates of emotional intelligence and psychometric intelligence: evidence for gender-based stereotypes. *Journal of Social Psychology*, **144**, 149–162.

Vidal Rodeiro, C.L. and Bell, J.F. (2007). Factors affecting examination success at A-level. *Research Matters: A Cambridge Assessment Publication*, **3**, 14–19.

Wang, M. C., Haertel, G. D. & Walberg, H. J. (1993). Toward a Knowledge Base for School Learning. *Review of Educational Research*, **63**, 3, 249–294.

Zins, J.E., Weissberg, R.P., Wang, M.C. & Walberg, H.J. (2004). *Building academic success on social and emotional learning: What does the research say?* New York: Teachers College Press.

Assessment instruments over time

Gill Elliott, Milja Curcin, Nat Johnson, Tom Bramley, Jo Ireland, Tim Gill and Beth Black Research Division

Introduction

As Cambridge Assessment celebrated its 150th anniversary in 2008 members of the Evaluation & Psychometrics Team looked back at question papers over the years. Details of the question papers and examples of questions were used to illustrate the development of seven different subjects. In each case the following research questions were addressed:

- Has the assessment structure altered over time?
- Have the emphases on different topic areas changed over the years?

The seven subjects studied were:

Mathematics	Physics	Geography	Art
French	Cookery	English Literature	

Background

In the 150 years since Cambridge Assessment/University of Cambridge Local Examination Syndicate has been in existence, there have been a great many educational and social changes affecting students, teachers and assessments. This project sought to describe some of these changes and to illustrate them through changes in question papers. The project was a departure from the usual qualitative and quantitative methods used by the Evaluation Team, and instead took the form of a semi-structured investigation of the development of a number of subjects through the questions presented in the written examination papers.

These studies cannot be used to provide a commentary on *standards* over time, for several reasons:

- First, they do not contain sufficient salient information about the mark schemes, the curriculum and the exact nature of the work produced in response to the questions (scripts). Without *all* of these pieces of information, most of which no longer exist, comparisons about whether a particular era is 'better' simply cannot be made.
- Secondly, examination questions have changed over the years. For example, advances in technology have made it possible to routinely calculate statistics about questions (e.g. facility values) which can provide question writers with important feedback about the performance of that question. Additionally, much development has occurred around question writing and question writer training. Older questions which may seem difficult to 21st century readers may have been difficult for reasons which

would nowadays be challenged on the grounds of fairness or validity. Finally, the regulation and oversight of all Awarding Bodies has changed beyond recognition in 150 years. Therefore, simplistically comparing questions from one era with another as evidence of changes in standards over time is flawed.

- Thirdly, the nature of the cohort has altered over the years and examination questions do not show this. So for example, the candidates sitting a School Certificate examination in 1907 might have been only a tiny proportion of the 16-year-old population, whereas the vast majority of 16-year-olds enter for GCSEs in the current context. As a consequence the level of accessibility of the questions differs – modern questions must be worded in such a way that all students being targeted can make some attempt at answering. The target candidature of past questions (particularly those from the earliest years sampled) was undoubtedly very different.

However, studies such as these can be used to illustrate the vast changes that have occurred, and the examples which follow show a small selection of the findings in each subject. These were presented as a poster at the 34th International Association for Educational Assessment (IAEA) Annual Conference which was hosted in Cambridge from 7–12 September 2008 by Cambridge Assessment, as part of the celebrations for its 150th anniversary.

The studies looked at the way in which papers were structured over the years, as shown in these examples from the **Physics** study (Table 1).

The two key themes which have been identified across many of the subjects include the increase in the number of questions relating to real-world contexts, and the greater amount of choice available to candidates, both in terms of the different options within assessments and the methods by which they may display their skills.

Increasing use of real-world contexts can be illustrated from the study into **Mathematics**, where it was interesting to note that as early as 1957 one of the regulations sections stated that some of the questions might be set on the application of certain arithmetical processes to problems of everyday life in the home and the community. This appears to be one of the early explicit statements indicating a trend that became prevalent in testing all topic areas of mathematics in the GCSE Mathematics papers, although it was present even in the 19th century papers to some extent, especially in the area of Arithmetic.

An example of a question from the 1997 GCSE Mathematics assessment:

Mrs McKenzie bought a large box of bags of crisps for her family. She told the children that the box should last 3 weeks if they ate 12 bags per week between them.

(i) How many weeks should the box last if the children eat 9 bags per week between them?

If the children eat n bags per week between them, the box will last W weeks.

(ii) Write down a formula which connects W and n.

The studies investigated how topic areas within subjects have altered over the years. In this example from the **Geography** study (Figure 1), physical geography, human geography and geographical skills have featured since early days, but economic and environmental geography are more recent elements of the assessment.

Table 1

Year	*Paper*	*Time*	*Rubric*	*Example question*
1927	Physics I	2hrs	Not more than **six** questions are to be attempted.	Explain the phenomenon of dew, and discuss the conditions which favour its formation. How is the dew point determined, and how can the relative humidity of the atmosphere be calculated when the dew point is known?
1957	Physics Ordinary Level Theoretical Paper	2½ hrs	Answer **all** the questions in Part I and **five** questions from Part II including at least **one** question from each of the Sections A, B, C.	{From Part I}: What is the freezing-point of water on the Fahrenheit scale? Express, in °C, a temperature which is 45 degrees below the freezing-point of water on the Fahrenheit scale.
2007	1982/4 Science: Physics extension option A Paper 4 Higher Tier	45mins	Wide range of mark totals per question	This question is about generating electricity. In 2005 the Prime Minister, Tony Blair, called for a 'National Debate' on nuclear power, climate change, and renewable energy sources. (a) Explain what is meant by a **renewable energy source**. [2] (b) More nuclear power stations could be built. (i) Suggest **two** arguments for building more nuclear power stations. [2] (ii) Suggest **two** reasons **against** building more nuclear power stations. [2]

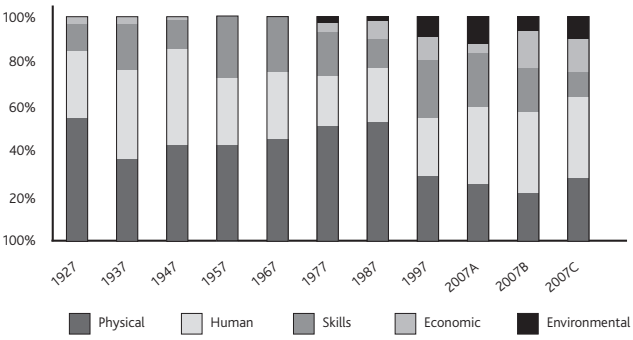


Figure 1: Geography: Summary of topic areas over time

In some instances practical considerations have affected the practice of assessment.

For example, **Artwork** (Figure 2) used to be necessarily restricted by weight and size, because the work was sent to Cambridge and displayed in the Craft Hall at 1 Hills Road for marking.

"Pieces of pottery must not exceed 12 ins. in any dimension, nor exceed 7 lbs. in weight. Pieces of sculpture or carving must not exceed three feet in any dimension nor exceed 20 lbs. in weight."

1977 and 1987 Art specification

Now that schools themselves display candidates' work and examiners make visits to the schools, students' artwork is not limited in this way.

Finally, the studies have enabled analysis of the skills required by candidates at different points in time in specific subjects. In **English**

Photograph by Peter Askem. Cambridge Assessment Archives Ref: M/P 5/8

Figure 2: Artwork in 1 Hills Road for marking

Literature every question paper between 1877 and 1937 inclusive (Table 2) required candidates to quote verbatim from memory fairly substantial sections of the prescribed text. Earlier question papers used to require candidates to know the precise meaning, usage and etymology of words in the texts, and on occasion, questions would require candidates to quote a line in which a particular word appeared. Later question papers gave more emphasis to discussing overall meaning or themes of a text and describing or analysing the candidate's own response to a passage or character. A particularly common feature of later papers asked candidates to imaginatively play the role of a character in the text.

Summary

The research proved a very interesting means of investigating the development of individual subjects. Naturally the method used – sampling question papers from every tenth year- has some limitations. It is, for example, possible that short-lived topics or question paper structures have escaped our attention altogether. Also the researchers are unable to state for certain exactly when a particular change occurred – the research shows merely the first sampled year when such changes were seen.

However, many interesting details have emerged from every subject studied and two themes were repeated across many of the subjects. These were an increasing emphasis upon real-world contexts for questions in more recent years, and an increasing choice of topic areas and question/component options available to candidates.

For full reports in each of the seven subjects, please contact Gill Elliott, Assessment Research and Development Division, Cambridge Assessment, 1 Regent St, Cambridge, CB2 1GG. Email: elliott.g@cambridgeassessment.org.uk.

Table 2: Skills tested over time in English Literature

Skill	*1877*	*1887*	*1897*	*1907*	*1917*	*1927*	*1937*	*1947*	*1957*	*1967*	*1977*	*1987*	*1997*	*2007*
grammatical analysis	✔		✔											
etymology	✔	✔	✔											
textual analysis	✔	✔	✔	✔										
scan (divide into metrics), knowledge of poetic/linguistic form (pentameter)		✔	✔	✔		✔	✔							
knowledge of author's life (external to text)			✔	✔		✔								
produce quotations verbatim	✔	✔	✔	✔	✔	✔	✔							
knowledge of literary, dramatic or poetic terms, concepts and mechanisms	✔	✔	✔	✔	✔	✔	✔	✔	✔					
translate text into contemporary prose retaining exact meaning	✔	✔				✔	✔		✔	✔				
comparison of text with factual information/external point of reference			✔	✔	✔	✔	✔				✔			
explain meaning of (extended) text (expound)	✔		✔	✔	✔	✔	✔		✔	✔		✔		
exact context of quote/excerpt	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔		✔		
overall evaluation of play/text/poem		✔	✔				✔	✔		✔	✔	✔		
give an account of a scene/sequence of events/story strand/poem	✔		✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	
character analysis/development including comparison of characters	✔		✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
thematic analysis/overall theme					✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
qualities of the writing of the poem/text						✔	✔	✔	✔	✔	✔		✔	✔
views or attitudes of the author as detected from the text.						✔	✔	✔	✔		✔	✔	✔	
appreciation of text/describing impact upon self/reader response						✔	✔	✔	✔	✔	✔	✔	✔	
significance (importance) of a feature or scene of text							✔	✔				✔	✔	
understanding of staging of play/dramatic impact							✔		✔				✔	
reading comprehension of text (within 'extract based questions')											✔	✔	✔	✔
relating to characters											✔	✔		
imaginative writing – role playing a character in the text											✔	✔	✔	✔
unseen poem											✔	✔		

All the right letters – just not necessarily in the right order. Spelling errors in a sample of GCSE English scripts

Gill Elliott and Nat Johnson Research Division

This article is based on a paper presented at the British Educational Research Association Conference in Edinburgh in September 2008.

Abstract

For the past ten years, Cambridge Assessment has been running a series of investigations into features of GCSE English candidates' writing – the Aspects of Writing study (Massey *et al*., 1996; Massey *et al*., 2005). The studies have sampled a fragment of writing taken from the narrative writing of thirty boys and thirty girls at every grade at GCSE. Features investigated have included the correct and incorrect use of various forms of punctuation, sophistication of vocabulary, non-standard English, sentence types and the frequency of spelling errors. This article provides a more detailed analysis of the nature of the spelling errors identified in the sample of work obtained for the Aspects of Writing project from unit 3 (Literary Heritage and Imaginative Writing) of the 2004 OCR GCSE examination in English. Are there certain types of spelling error which occur more frequently than others? Do particular words occur over and over again? How many errors relate to well-known spelling rules, such as 'i before e except after c'?

Literacy has enjoyed a high profile since 1994 and has been promoted in schools through the introduction of the National Literacy Strategy (NLS). It was unlikely that the 2004 GCSE cohort (the 'population' from whom our writing sample came) was fully exposed to the NLS. This is because many primary schools introduced the NLS from the bottom up, or at least did not implement it for this cohort (in their final year of primary education in the first year of the NLS) on the basis that it would get in the way of preparation for key stage 2 (KS2) national tests (Beverton and English, 2000). This notwithstanding, Beverton and English noted that, in contrast to previous years, grammar was being taught every day and that all teaching staff in the schools observed had a greater awareness of literacy as a subject in its own right. Therefore, the performance of this cohort in spelling is likely to reflect some of the benefits of the NLS.

The study used a stratified random sample of writing taken from a narrative writing task. The only suitable question was found on a paper which formed an alternative to coursework; a question which asked candidates to imagine, rather than to inform, explain, describe, comment, argue or persuade. This option was taken by only 8.3% of candidates – but these amounted to over 5500 candidates from a wide range of schools. The sample was stratified by grade so the fact that this paper was a minority option should be incidental, as the calibre of a candidate achieving a particular grade should be comparable regardless of the route taken through the syllabus. Whilst the possibility existed that schools choosing the examination option might reflect systematic variations in curricular values, comparison of the examination option schools with the entry as a whole did not suggest that the former were unusually socially or educationally selective. The proportions of independent and selective schools as compared with comprehensives and others were the same for the sample as in the overall entry for this English specification.

Spelling errors were identified in the sampled writing by two researchers, working first separately, and then as a team. Each researcher first went through the printed versions of the script samples identifying and counting spelling errors. The two lists of errors and counts were then compared, again grade by grade, and any discrepancies identified and discussed.

The study identified 345 spelling errors in 11,730 words written, and these were reported in Massey *et al*. (2005), with a comparison by grade with samples of writing from 1980, 1993 and 1994. It was shown that a considerable decline in spelling in the early 1990s (compared with 1980) had been halted, and at the lower grades, improved.

Since then, we have conducted a detailed analysis of the 345 misspelled words to see if there is evidence of particular types of error. Each misspelling has been categorised, and five broad types of error identified. These are:

i. sound-based errors,
ii. rules-based errors,
iii. errors of commission, omission and transposition,
iv. writing errors and
v. multiple errors.

This article will present a detailed examination of the misspellings and the process of developing the categorisation system used. A number of words – *woman, were, where, watch(ing), too* and the homophones *there/their* and *knew/new* are identified as being the most frequently misspelled words. Implications for the findings upon teaching and literacy policy are discussed.

Background

The way in which children learn to spell is linked closely to learning to read, and with other elements of learning to write. Westwood (2008) reviewed the literature from 1995 to 2007 pertaining to the strategies used to teach children to read in English in Australia and Great Britain and Wanzek *et al*. (2006) published a review of a large number of intervention studies carried out between 1995 and 2003.

A number of authors have looked at stages by which a child learns to spell. Ehri (1994) identified a 'logographic' stage, whereby a child deduces meaning from the appearance of the words. Later stages include the ability to match letters to speech sounds (Henderson, 1990) and use

these to decode words (read) or to generate their own words (spell). Moats (1995) suggests that a phonetic spelling stage is then attained, with children following a 'one letter spells one sound' strategy. This is the point at which spelling can deviate from conventional 'correct' spellings, especially in English where sound rules do not necessarily match letter rules. At this point the successful speller must memorise specific rules such as grammatical endings, and different words which sound the same but are spelt differently. A study carried out between 1995 and 1998 by the Centre for Language in Primary Education (O'Sullivan and Thomas, 2000) collected data from London primary schools and investigated the teaching and learning of spelling throughout the primary years. Amongst other findings the study reported that it is helpful for teachers to study the mistakes made by individual spellers, in order to assess whether the mistakes they are making are phonetic or visual.

In the UK there have been two main methods of teaching a child to read – synthetic phonics, where children are taught letter sounds before being introduced to whole words (Auger and Briggs, 1992), and analytic phonics, where whole words are introduced from the start. Johnston and Watson (2003, 2004, 2005) have suggested that the reading and spelling skills developed by children taught to read using synthetic phonics are very good.

A number of frameworks already exist which incorporate categories of spelling error. QCA (1999) mentions errors due to unstressed vowels, long 'e', omission of single letters, confusion of consonants and homophones. Homophones are also a feature studied by Hepburn (1991) along with doubling and singling of consonants, articulation, and errors related to inflectional and derivational morphemes. Finally, Mudd (1994) discusses reasonable phonic alternatives – in other words plausible alternative spellings.

Method

The sample of writing from which the spelling errors were identified consisted of the fourth sentence[1] of question 1 (an extended narrative piece of writing) as written by the candidate, and was taken from the scripts of thirty boys and thirty girls at each grade. Where there were insufficient suitable scripts available additional sentences were taken from available scripts. The sentences sampled were keyed into Word™ by a temporary member of staff, preserving all errors of punctuation and spelling. Careful checking was undertaken to ensure that the keying, including errors, had been accurately undertaken. Counts of the numbers of words were then obtained from Word™ software.

Table 1 shows the number of words which were sampled at each grade.

Table 1: Number of words sampled at each grade

Grade	*A**	*A*	*B*	*C*	*D*	*E*	*F*	*G*
Number of words	1238	1082	1303	1208	1567	1734	1739	1859

Spelling errors were identified by two members of staff, working first separately, and then as a team. Each person first went through the printed versions of the script samples, grade by grade, identifying and counting spelling errors. The two lists of errors and counts were then compared, again grade by grade, and any discrepancies identified and discussed. At any stage it was also possible to inspect the handwritten scripts to verify the exact marks placed on the paper by the candidate. The benefit of the doubt was given in any case where there was ambiguity, which usually arose as a consequence of either poor handwriting, or poor spacing technique. In some cases it was necessary to look elsewhere in the candidate's script for examples of particular letters or letter combinations, or to look at the spacing between other words to see whether the presence or absence of spacing appeared to be deliberate on the part of the candidate.

1 Everything which appeared between the third and fourth full stop.

Results

Overall numbers of spelling errors

The study identified 345 errors in 11,730 words written. Therefore, 97.1% of words were correctly spelled.

Figure 1 shows the overall numbers of spelling errors by grade. As expected, the number of errors increases by descending grade. Given that spelling errors are one of the (admittedly many) criteria for judging English writing, it would be unexpected if they did not. Figure 2 shows the same data as a percentage of the total number of words, thus adjusting the bars for the number of words written in total (candidates at different grades wrote different numbers of words, and as every word written presents an opportunity for a spelling error, variability in the total number of words might influence the pattern of results). In fact the adjusted graph remains very similar to the raw data.

This paper provides detailed analysis of all the errors to see if it there is evidence for particular types of error. Appendix 1 gives the entire list of words which were spelled wrongly, arranged in alphabetical order.

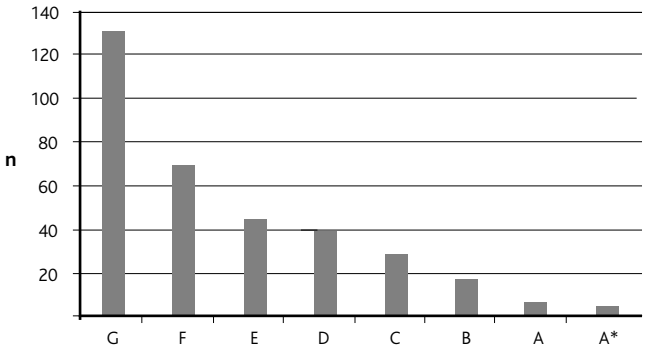


Figure 1: Number of spelling errors by grade

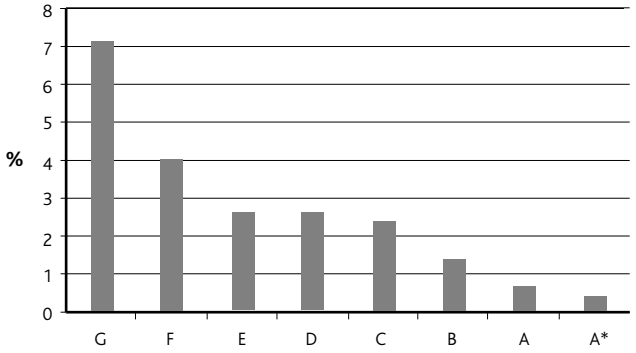


Figure 2: Rate of spelling errors by grade

Frequently occurring misspelt words

A few words occur more frequently than others. Words which appear in the list more than twice are listed in Table 2, along with the frequency of their occurrence, a list of each misspelling and a list of the grades at which the misspellings occur. (The misspellings and corresponding grades are given in the same order, to enable the reader to identify which particular misspelling occurs at which grade.)

Table 2: Frequently occurring misspelt words

Word	*Freq.*	*Misspellings*	*Grade*
before	3	*befor, befor, be for*	GGG
finally	4	*finaly, finily, finaly, finaly*	BGGG
here	3	*he, hear, hire,*	EFG
knew	5	*new, new, new, new, new*	DEEFF
their	3	*ther, there, thire,*	FFF
there	10	*ther, their, their, their, the, their, their, their, their, ther*	BCEEFFGGGG
they	6	*thay, thay, thay, thay, thay, thay*	GGGGGG
too	4	*to, to, to, to*	DEFF
towards	3	*to-wards, to words, to words*	EGG
until	3	*untill, untill, untill*	CCF
watch(ing)	4	*wach, waching, waching, waching*	EFGG
went	3	*when, when, whent*	GGG
were	5	*where, where, where, where, where*	CDDDG
where	4	*were, were, were, were*	DDDF
woman	11	*women, women, women, women, women, women, women, women, women, woneman, women,*	BCDDEEEEFFG
you	3	u, yo, yoy	GGG

Seven of these words – *here, their, there, too, were, where, you* – appear in published Key Stage 1 lists and *before, knew, until, watch, woman* all appear in Key Stage 2 lists.

Although *women* for *woman* is the single most frequently occurring mistake with ten instances (and occurs at every grade from B downwards), the *their/there* homophone is a close second, with eight occurrences, seven of which are *there* for *their*.

Misspellings by type

The misspellings presented by candidates have been grouped into broad related categories of error. Categories were derived via a process of grouping together similar error patterns, and are shown in Figure 3. As far as possible the 'types' of error were kept as simple as possible, in the spirit of the original Aspects of Writing (the generic name given to the series of reports produced by Cambridge Assessment, and its predecessor, UCLES) research. This resulted in the following categories:

- Sound-based error – homophones, incorrect consonant, e for y, vowel sound error, morpheme error.
- Rules based error – doubling/singling, text-speak.
- Omission, commission and transposition – single or paired letters added, omitted or transposed.
- Writing error – spacing, end of word missing.
- Multiple errors.

Where a misspelling might fall into several categories (i.e. *accross*, which is both a doubling error and the insertion of an additional letter) the most obvious/most precise error type was allocated; in this case, doubling).

Discussion of error types

Sound-based

Homophones form the first category of error types. 34 of the 345 errors (9.8%) were of this type. The *there/their, know/no* and *knew/new* confusions accounted for nearly half of these. These errors have already been discussed in the section on frequently occurring misspelt words.

Fifteen errors consisted of the transposition of a single wrong consonant. Many of these were phonetically plausible spellings; however, there were instances of a 'k' at the end of *–ing*, instead of the 'g', and of 't' replacing 'd' in *–ed* endings. These were potentially due to articulation error, resulting in spelling error. Two errors involved the transposition of a vowel for a consonant – in both cases 'e' for 'y'.

Fifty-two errors related to the vowel sound. Again (or *agen* according to one such candidate), most of these were phonetically plausible spellings. Nonetheless, many of these words are to be found on the lists of spellings at KS1 and KS2 – e.g. *hospital, heard, some, doctor, they*.

Rules-based

Doubling/singling errors

There were 13 doubling errors and 22 singling errors, together accounting for 10% of all errors. Only one of the errors (*aclimatised*) was an example of an affix error.

Suffix errors

There were 24 suffix errors (7% of the total), of which a very high proportion involved adding *–y* or *–ly* to a word or involved the 'y' to 'i' rule (changing a *y* to an *i* before adding *–ed* (e.g. *replyed*).

Two errors were 'text' (mobile phone/computer text messaging) influenced. Once again these are phonetically plausible alternatives to conventional spelling and are intentionally used in defiance of 'conventional' spelling rules during text messaging. The very small number of these errors was remarked upon in the original report, and it is pleasing to see that candidates seem by and large to be aware that they must not use such devices in a written English examination, however much they are used in social contexts.

Omission/commission of single letter and transposition

Forty-nine errors consisted of the omission of a single letter, whilst thirty-four were the insertion of a single letter. In some cases these were clearly the result of idiosyncratic spellings – notably silent letters. In other cases, the error perhaps owes more to carelessness.

Only ten errors were a straight reversal of two letters, and just one of these related to the 'ie/ei' rule.

Writing errors

Two types of error have been categorised as 'writing' errors. These are errors of spacing – writing two words as one or vice versa, and missing the last letter from a word. In several instances there is evidence from the scripts that candidates did know the correct spelling in the case of the latter category, but had left off the final letter in haste.

Multiple errors

These errors form the arguably most striking type of mistake, and have most effect upon the appearance of the word. First are those misspellings which seem to be made up from two separate errors. For example:

Figure 3: Misspellings by type

Sound-based

Homophones

greatful ruff
hear stairing
here their (x6)
layed there (x2)
lent to (too) (x4)
new (x5) to (two)
no (x2) warn
past weather
piecefully who's
road your

Single consonant confused with another single consonant

ang pup
edje reseption
glanze somethink
looket startet
nothink surport
pankakes trappling
pass warking
pud

e for y

empte lade

Phoneme- grapheme mismatch

agen practicle
clame quiot
comfatable re-esuring
cud saed (x2)
deap sead
devestating screeming
docters secutary
egere suffercated
examaning sume
extreamly survay
frale thay (x6)
hospitel tomarto
hourse trough (true)
hurd tumer
nely (x2) uncomfertable
nieve weerdos
paitients women (x10)
parshly wonted

Rules-based

Suffix rules

ly

definatly slightley
funnyly slowley
highley unnaturely (x2)
luckly

drop e before adding -ing

closeing hopeing
comeing (x4) stareing

y

angery inevitabely
angrey panicy
babys scarey

y to *i*

replyed tryed

ed

answerd offerd

Doubled consonant where should be single

accross ponny
allways pressumably
harrassed quiettly
normall ridding
openned unoccuppied
untill (x3)

Single consonant where should be doubled

acident ofering
aclimatised penciled
asortment poped
atempt siting (x2)
caled spliting
comotion stifly
embarassment stoped (x2)
finaly (x3) sufering
gona tanoy

Text influenced

thanx
u

Omission, commission, & transposition

Single letter omitted

assiting plasic
attemted plesant
belive (x2) quit
complant quitly
coner relised
consious scrunced
contined set
crowed (x2) skatebord
denist stared
drumsicks (started)
easly stiped
emty stroger
enviroment studing
everone subconsiously
exept suprise
exusing suprisingly
fustrating tak
grove tepted
is (his) the (they)
newpaper throt
nuber wach
ofering waching (x2)
overwelming wat
pachy were (where)

Unnecessary letter inserted

alls otheir
anixiously plance
anouther propbable
diden't site (sit)
disrupte smocking
doupe stat
hand (and) throught
has (as) tould
heared verey
hoppe watiching
markers whant
minde whas
minuites whent (x3)
off (of) where (x5)

Extra syllable inserted

partening woneman

2 letters reversed

brian minuets
ect recieve
frist retruned
gentelman thier
minuet wrinkeld

Writing

Spacing/writing two words as one or vice versa

alot to-wards
alright a nother

End of word missing

befor (x2) of
ever on
feminin the (there) (x2)
gonn (gonna) ther (there)
he (her) though
he (here) tong (tongue)
I (it) use (used) (x2)
imagin yo
nam

Multiple

Two 'simple' errors

abound (about) oader
angshuse poedem
babal reapted
behide remmeberd
be for sopose
costrophobic sopted
diesese stir (stare)
diside sufficate
enbarrased sumbleing
finily to words
formiler toke
glome tort
gourges unaturaly
imediatly when (went)
impaitientley (x2)
manegd wittnes
nieghbor

Part of word missing/ severely misspelled

apoched
appment
blacks (blackouts)
canures (cancerous)
he's (here's)
imaging (x2)
prespetion
prest (pressed)
pust (pushed)
registed
scowered
themsefs
trould (trouble)

3 or more mistakes

alla (all of)
handon (handsome)
immeiadtley
solisters
thire (their)

Extreme phonetic errors

ant shaght (anxious)
asct (asked)
corried door (corridor)
faunt (thought)
hast (asked)
nufse (nervous)
or wright (all right)
torck (talk)

impaitientley consists of two separate inserted letters; *impa(i)tientl(e)y*

nieghbor consists of a transposition and an omitted letter (in UK spelling);*n(ie)ghbo(u)r*

Second are those errors where a whole part of a word is either missing or severely misspelled. The third category within this group contains those few words with three or more individual mistakes, and it was one of the misspellings – *immeiadtley* – which prompted the title of this article – all the right letters, just not necessarily in the right order. Finally, there are a group of words which bear little physical resemblance to their correct spellings, yet have clear phonetic links with them. These are referred to as extreme phonetic errors. It is possible that this latter category may be related to the very specific types of error made by people with dyslexia, but further discussion of this is beyond the scope of the present article.

Discussion

This article has attempted to categorise spelling errors made by students in their GCSE English examination in 2004 into various categories. The purpose of the research was to establish whether certain spelling errors – or certain categories of error – are particularly common, and how they relate to spelling conventions, as taught within schools.

The study has identified five categories of spelling error which further subdivide into sixteen subsections. The categories were derived from the errors observed, rather than from existing categories, so there may be other groups of spelling error which have not been discussed here, simply because they were not encountered. In general, most misspellings fall into the first three categories: sound-based error, rules-based error and errors of omission, commission and transposition. The first two of these categories contain many misspellings that are undoubtedly very familiar to teachers. However, there are no particular sub-categories that are particularly prone to more errors in our sample than others. English is a language which has more than its fair share of idiosyncratic spellings and complex spelling rules. Not surprisingly, many of these errors are connected with those. However, within the category of a single additional letter, there were a number of examples of an unnecessary silent 'h' – *where (were)*, *whant, whas*, which are worthy of comment. The category of omission, commission and transposition is more difficult to interpret. It is quite possible that many of these errors occurred as a result of the examination conditions under which candidates were writing, combined with, perhaps, a lack of effective proof-reading of their final piece. The sub-category of writing errors, where the ends of words are missing, could in some cases be due to the same issues. However, the spacing of two words as one, or vice versa, is almost certainly due to candidates' perceptions of those words. Finally, the category of multiple errors produces words which look least like conventional spellings. Interestingly, two simple errors can produce a word that is almost unrecognisable, and it is important to be able to decode these errors for what they are, rather than simply seeing a very distorted word.

Fifteen individual words were identified as occurring with relatively high frequency. In particular, two of these were seen far more often than others. They were the *there/their* homophone, which has been known to be problematic since time immemorial, and *women* for *woman* (not vice versa). *Knew/new* and *know/now* also occurred with relative frequency, but again, this is unlikely to surprise the teaching profession.

A major limitation to the data presented here is the fact that there is no control over which words candidates choose to use. Therefore the study is not a 'fixed' spelling test, and cannot be generalised in the same way as reports of spelling tests. A word spelt wrongly just once does not mean that 479 students can spell it, simply that they did not necessarily try. It would be possible to investigate correctly spelt words to give the other side of the picture, but that would be an enormous task.

There is clearly no single over-riding type of error which is made by the group of GCSE students from whom we have sampled. Those errors that are made are varied, and although it is disconcerting to note the number of most frequently occurring errors which are taught at Key Stage 1, it is, on the other hand, heartening to see how few (relatively speaking) errors are made, when you consider the number of words written overall, especially given that the text was written under examination conditions with no access to dictionaries.

References

Augur, J. & Briggs, S. (1992). *Hickey Multi-Sensory Language Course*. London: Whurr Publishers Ltd.

Beverton, S. & English, E. (2000). How are schools implementing the National Literacy Strategy? *Curriculum*, **21**, 2, 98–107.

Ehri, L. C. (1994). Development of the ability to read words: update. In: R. Ruddell, M. Ruddell & H. Singer (Eds.), *Theoretical Models and Process of Reading*. Newark, Del.: International Reading Association.

Henderson, E. (1990). *Teaching Spelling*. Boston: Houghton Mifflin.

Hepburn, J. (1991). Spelling categories and strategies. *Reading*, April 1991.

Johnston, R. S. & Watson, J. (2003). *Accelerating Reading and Spelling with Synthetic Phonics: A five year follow up*. *Insight 4*. Edinburgh: Scottish Executive Education Department.

Johnston, R. S. & Watson, J. (2004). Accelerating the development of reading, spelling and phonemic awareness. *Reading and Writing*, **7**, 4, 327–357.

Johnston, R. S. & Watson, J. (2005). *A seven year study of the effects of synthetic phonics teaching on reading and spelling achievement*. *Insight 17*. Edinburgh: Scottish Executive Education Department.

Massey, A. J. & Elliott, G. L. (1996). Aspects of Writing in 16+ English examinations between 1980 and 1994. Occasional Research Paper 1. University of Cambridge Local Examinations Syndicate.

Massey, A. J., Elliott, G. L. & Johnson, N. K. (2005). Variations in aspects of writing in 16+ English examinations between 1980 and 2004: Vocabulary, Spelling, Punctuation, Sentence Structure, Non-Standard English. *Research Matters: A Cambridge Assessment Publication*. Special Issue, November 2005.

Moats, L. C. (1995). *Spelling: Development, Disability and Instruction*. Baltimore: York Press.

Mudd, N. (1994). *Effective Spelling: A practical guide for teachers*. London: Hodder & Stoughton.

O'Sullivan, O. & Thomas, A. (2000). *Understanding Spelling*. London: Routledge.

Qualifications and Curriculum Authority (1999). *Technical Accuracy in written English: Research findings*. London: QCA.

Wanzek, J., Vaughn, S., Wexler, J., Swanson, E. A., Edmonds, M. & Kim, A.H. (2006). A synthesis of spelling and reading interventions and their effects on the spelling outcomes of students with LD. *Journal of Learning Disabilities*, **39**, 6, 528–543.

Westwood, P. (2008). Revisiting issues in spelling instruction: A literature review 1995–2007. *Special Education Perspectives*, **17**, 1, 33–48.

Appendix 1: Alphabetic list of words which were spelled wrongly in a sample of GCSE English writing

A
about
accident
acclimatised
across
again
all of
all
always
a lot
all right
and
angry
another
anxious
anxiously
answered
approached
appointment
as
asked
assortment
assisting
attempt
attempted

B
babble
babies
before
behind
believe
blackouts
brain

C
called
cancerous
claim
closing
coming
comfortable
commotion
complaint
corner
conscious
continued
corridor
claustrophobic
crowded
could

D
deep
decide
definitely
dentist
devastating
didn't
disease
disrupt
doctors
dope
drumsticks

E
eager
easily
edge
embarrassed
embarrassment
empty
environment
etc
every
everyone
examining
except
excusing
extremely

F
face
familiar
feminine
found
finally
first
frail
frustrating
funnily

G
gentleman
glance
gloomy
gonna
gorgeous
grateful
groove

H
had
handsome
harassed
her
hear
heard
here
here's
highly
his
hope
hoping
hospital
horse

I
imagine
imagining
immediately
impatiently
inevitably
it

J
K
knew
know

L
lady
laid
leant
looked
luckily

M
makers
managed
mind
minute
minutes

N
naive
name
nearly
neighbour
nervous
newspaper
normal
nothing
number

O
odour
of
off
offered
offering
one
opened
other
overwhelming

P
pancakes
panicky
partially
parting
patchy
patients
passed
past
peacefully
pencilled
pleasant
place
plastic
podium
pony
popped
practical
prescription
pressed
presumably
probably
pub
pushed

Q
quiet
quietly
quite

R
realised
reassuring
reception
receive
registered
remembered
repeated
replied
returned
riding
rode
rough

S
said
sat
scary
scoured
screaming
scrunched
seat
secretary
sit
sitting
skateboard
slightly
slowly
smoking
solicitors
some
something
splitting
spotted
stare
staring
started
stiffly
stopped
striped
stronger
studying
stumbling
subconsciously
suffering
suffocate
suffocated
support
suppose
surprise
surprisingly
survey

T
talk
tannoy
tempted
thanks
their
themselves
there
they
thought
throat
too
told
took
tomato
tongue
towards
trampling
tried
trouble
true
tumour
two

U
uncomfortable
unnaturally
unoccupied
until
used

V
very

W
walking
want
wanted
was
watch
watching
weirdoes
went
were
what
whether
whose
where
witness
woman
worn
wrinkled

X
Y
you
you're

Statistical Reports

The Statistics Team Research Division

The Statistics Reports Series is based on the annual national-level examination databases for pupils in England. The objective of these reports is to provide statistical summaries of the information contained in these, such as pupil attainment, pupil subject uptake and subject provision in schools. Some reports will consider patterns over time, particularly if there has been a relevant change in the system. These reports will be produced at a rate of around two or three a year and are available in .pdf format on the Cambridge Assessment website: http://www.cambridgeassessment.org.uk/ca/Our_Services/Research/Statistical_Reports

As of December 2008, the following reports are available:

Statistics Report Series No.1: Provision of GCE A-level subjects

Statistics Report Series No.2: Provision of GCSE subjects

Statistics Report Series No.3: Uptake of GCE A-level subjects in England 2001–2005

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

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

Statistics Report Series No.6: Numbers of A-level examinations taken by candidates in England 2006 and the percentages attaining 3 or more A grades

Statistics Report Series No.7: The relationship between A-level grade and GCSE grade by subject

Statistics Report Series No.8: Uptake of GCSE AS level subjects in England 2001–2007.

Statistics Report Series No.9: Numbers achieving 3 A grades in specific A Level combinations by school type and LEA

De-mystifying the role of the uniform mark in assessment practice: concepts, confusions and challenges

Elizabeth Gray OCR **and Stuart Shaw** CIE

Introduction

The search for an adequate conceptualisation of the Uniform Mark Scale (UMS) is a challenging one and it is clear that there is a need to broaden current discussions of the issues involved. This article marks an attempt to demystify the UMS; its conception and operation. Although the article assumes a basic appreciation of the terminology and processes associated with the examination system, it explicates through a number of case study scenarios, the contexts in which it is appropriate to employ UMS, describes any necessary computations arising from different specifications and assessment scenarios, and addresses some of the potential challenges posed by the calculation of grades for unitised specifications. A specification here refers to a comprehensive description of a qualification and includes both obligatory and optional features: content, and any performance requirements. If a specification is unitised, the constituent units can be separately delivered, assessed and certificated.

Having a clear and well articulated position on the underlying theory of UMS is necessary to demonstrate transparency with regard to the estimation of aggregate performance on unitised assessments and to support any claims we wish to make about the reporting process.

It is hoped that the issues addressed here will make a positive contribution to the widening nature of the UMS debate (both within and beyond Cambridge Assessment) more generally, and of the understanding, operation and employment of UMS, in particular.

Underlying rationale for a Uniform Mark Scale

Educational assessments are currently delivered as either non-unitised specifications or as unitised ones. By non-unitised, we mean that candidates take the various components (which may be written papers, coursework or controlled assessment) that make up the specification in the same session or administration. Following any examiner or moderator scaling adjustments, the marks for candidates are aggregated to give a total mark for the entire specification: this defines the specification (also referred to as syllabus by Cambridge International Examinations, CIE) grade. The purpose of the grading process is to determine, the lowest mark for which the performance in the current administration can be deemed equivalent to that achieved by candidates at the lowest mark for the same grade for the last administration. Grading is undertaken for each key threshold on each component and each specification option. Generally the grading process attempts to involve comparisons with the standards set in previous sessions, but occasionally the process is one of standard setting for a new specification. Using the grade boundaries established by an awarding committee, a candidate's specification grades are subsequently determined from the total marks.

Unitised assessments, however, allow the candidate to take the unit assessments (the smallest part of the specification formally reported and certificated) on different occasions. Unitised assessments may exhibit variation in their respective levels of difficulty over time. Where this happens the grade boundaries for a January unit assessment, say, may be slightly different from those set for the corresponding May/June unit assessment. It is crucial, therefore, that a mechanism be implemented for mapping different marks awarded on different occasions onto some common scale such that the differing marks constitute the same value when aggregated to give an overall grade.

Issues relating to Aggregation

Aggregation is 'the process of combining (by summation or other agreed procedure) the marks or other units of credit awarded through an assessment scheme' (QCA Code of Practice, 2007, p. 65). Aggregation issues are a source of constant debate within the public examination area and Thomson (1992) provides a good description of the issues relating to methods which seek to combine raw marks (the marks originally awarded when assessed) of units achieved at different times and with different grade boundaries.

Potentially, there are a number of methods for combining raw marks of units achieved at different times. According to Thomson, many give rise to one of two types of anomaly. In Type I anomalies, two candidates with the same grade profile across four units receive different subject grades. For example, 'abbd' = B; and, 'abbd' = C. A special case of this is a candidate who obtains the same grade for all units, but obtains a different subject grade, for example, 'bbbb' = A. In Type II anomalies, two candidates with a different profile obtain the same grade. For example, 'abbc' = B; and, 'aabb' = B[1].

Different methods of aggregation give rise to different instances of these anomalies. Unless there is a very crude system of assigning a point to a grade, all methods will result in at least some Type II anomalies, and many in Type I. One of the reasons for the choice of uniform marks for aggregating unitised schemes is, therefore, that the instances of anomalies can be reduced if the conversion is suitably chosen (Thomson, ibid).

With the introduction of unitised schemes of assessment, GCE became wholly unitised in 2001/2002 although there were modular forms of general qualification assessments before then, and these add an additional complexity to the aggregation process because units may be taken within the duration of a course of study, not just terminally. In order to be fair to these candidates when a specification grade is calculated raw marks cannot be used. The reason is perhaps best illustrated by use of an example:

1 Clearly different grade profiles can lead to the same or different overall outcomes, some of which may be counter-intuitive.

Imagine a candidate takes a unit twice and achieves 72 raw marks in the first instance and 68 on re-sitting. On the first occasion the A boundary is set at 73 whilst on the second it is set at 67. In the case of the re-sit, the candidate gains a higher grade with a lower mark than in the first examination and the 'value' of raw marks is not the same for the two examinations.

An elementary approach for resolving this difficulty might be to award grades only to candidates on each unit. Unit grades would then be assigned numerical points (A*=9, A=8 ... U=1) and then a simple addition of points would provide a total for the specification. There are two distinct disadvantages with such a rudimentary method:

1. this approach would not discriminate between weak, adequate and strong performances within the same grade, in other words marks provide more detailed information than grades; also
2. problems would arise where units were unequally weighted. The weighting of an assessment is its overall contribution to the total or aggregate assessment. For example, if a unit is weighted at 35%, the unit accounts for 35% of the total assessment. In this case, a scaling factor would need to be applied to the raw marks in order to give them the appropriate weighting.

In order to obviate these shortcomings, a segmented linear scaling methodology is used. Such deficiencies are thus re-mediated by adoption of a procedure which utilises a common mark or standardised scale. A standardised mark is the result of a transformation of raw marks which provides a measure of relative standing in a group and allows comparisons of raw marks from different distributions (Davies *et al*. 1999, p. 186). A common scale[2] has the advantage of affording greater credit to candidates who have achieved higher marks within a grade and legislating for unequal unit weightings by setting a uniform mark scale for the unit which reflects its weighting in the specification.

In outline, raw marks are mapped on to a scale which takes into account the value of the raw mark. This scale is known as the uniform (or standardised) mark scale. Here, if the unit is worth 100 UMS then the A boundary is 80 UMS (using the usual GCE UMS). Taking the example introduced earlier: on occasion 1, the boundary of 73 raw marks would map to the UMS boundary of 80 and the candidate would get 79 uniform marks. On occasion 2, the boundary 67 raw marks would map to the UMS boundary of 80 uniform marks and the candidate would get 81 uniform marks. In this way the value of the raw mark, in terms of the grade it would earn, and the quality of that grade, are preserved.

The important point to note here is that uniform mark scales remain the same throughout the lifetime of the specification and, particularly, from one session to another. This means that the grade which a candidate receives and the position of the raw mark within the grade bandwidth (i.e. the marks between the two grade boundaries within which the raw mark sits) will always convert to the same uniform mark irrespective of the actual raw mark and the raw mark boundaries.

Thus a *uniform mark* is used when units of an assessment can be taken on different occasions during a course of study and is a mark on a standard scale which indicates a candidate's performance. A *uniform mark scale* is a means of achieving parity between alternative units in specifications and functions to effectively smooth out the small variations in the demand[3] of the assessment units sat by candidates during their GCE and GCSE studies. *Uniform mark boundaries* for unit and specification conversions remain the same for the lifetime of the specification.

It is a requirement of the QCA that aggregate marks from a unitised GCE or GCSE or staged tests should be computed on the basis of a UMS: 'Uniform marks for each unit must be calculated in such a way as to maintain the candidates' relative position between the raw grade boundaries. Each unit must be reported in uniform marks. Uniform marks for individual assessment units are added to generate a final grade for the qualification as a whole' (QCA Code of Practice, 2007, p. 56). The requirement for converting raw marks to uniform marks for the purposes of aggregation facilitates fairness of the specification outcomes.

The relationships between uniform marks and grades are shown in the relevant GCE and GCSE specifications and uniform marks and unit grade results are distributed to centres in the Cumulative Specification Results Report and to candidates in their Statement of Results.

We now turn our attention to how a uniform mark is calculated and in the computation process begin to appreciate some of the potential challenges unearthed by aggregation, highlighting some of the relative merits and de-merits of unitised schemes of assessment.

Conversion of raw to uniform marks

We have seen that a candidate's raw marks are mapped onto a scale which is invariant for the lifetime of the specification. The conversion of raw to uniform marks is dependent on the grade boundaries on the occasion when the raw mark was achieved.

The uniform mark scale will have been determined when the specification was originally accredited. So, for each unit, uniform mark maxima and grade boundaries are pre-set. In order to preserve the value of a raw mark, there should be a one-to-one, linear mapping of the uniform marks with the raw marks. In this instance, the boundaries are reasonably spaced and there are no issues relating to effects which manifest at the extremes of the mark distribution. This is illustrated in Figure 1 which assumes there are 100 raw marks. In this case, the GCSE A* boundary of 90 uniform marks will coincide with the raw mark boundary of 90 raw marks.

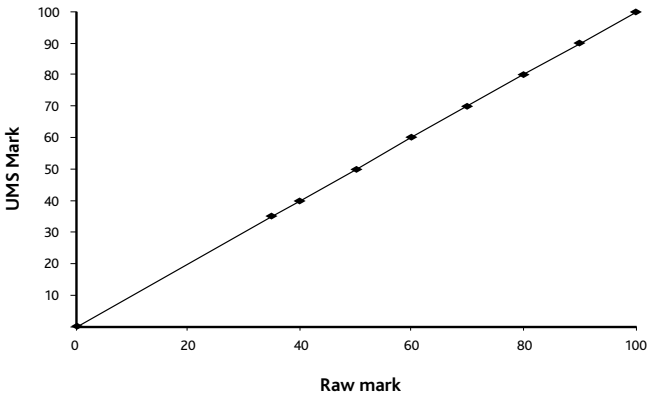


Figure 1: Ideal raw mark/uniform mark relationship

2 The simplest form of common scale would award one point per grade for equally weighted units which would not differentiate between candidates within a grade. It is rarely used in general qualifications for this reason.

3 Demand in this context is defined by awarders when setting grade boundaries. This is a judgement made in the presence of performance (candidates' work) and statistical evidence.

Figure 2 depicts another simple raw to uniform mark conversion which demonstrates linearity between grades A and E. However, above and below the two end points the conversion factor changes as is shown by the change in the slope of the line. This is because grades A and E are recommended at the grade award and intermediate grades are interpolated maintaining (to within a mark) the linear relationship of raw marks to uniform marks. However, unless A and E are chosen so that the mapping would continue the straight line between A and the maximum and E to zero then the line will consist of three segments. In fact, in Figure 2, raw marks above A and below E would be worth a smaller number of uniform marks than each individual raw mark between the A and E boundaries. The conversion factor from raw to uniform marks is smaller above the A boundary and below the E boundary than between the A and E boundaries.

Figure 2: A simple UMS conversion

In order to convert from a raw to a uniform mark:

1. Determine what grade the raw mark would indicate based on the raw mark boundaries for the unit.
2. Calculate the number of raw marks in the raw mark grade bandwidth.
3. Calculate the number of uniform marks in the uniform mark grade bandwidth. This number will be based on the weighting of the unit and the maximum UMS for the specification. All uniform marks between the top and bottom of the grade range will be the same, but may differ above the top and below the bottom grade.
4. Calculate a conversion factor (the number found in Step 3 divided by the number found in Step 2).
5. Find the difference between the lower raw mark grade boundary and the raw mark for conversion and multiply it by the conversion factor (from step 4).
6. Add the number found in Step 5 to the lower uniform mark grade boundary.

We can see how this process is operationalised in a live context by giving consideration to the following scenario:

Imagine a GCE candidate gains a raw mark of 49, the raw mark grade boundary for grade C is 47 and, the raw mark boundary for B is 53. Additionally, the UMS B boundary is 70 and the UMS C boundary is 60.

1. *The candidate has 49 raw marks which lie between the C and B boundaries so the candidate has a grade of C.*
2. *The raw mark B boundary is 53 and the raw mark C boundary is 47 giving 6 marks in the grade bandwidth.*
3. *The UMS B boundary is 70 and the C boundary 60, i.e. there are 10 uniform marks in the uniform mark grade band width (range).*
4. *The conversion factor is found by dividing the number of uniform marks in the range (10) by the number of marks in the grade bandwidth (6) i.e. 10/6= 1.7.*
5. *The candidate's raw mark is 49 and the lower raw mark grade boundary is 47. So the number of marks above the grade boundary is 49–47 = 2. The result of this calculation is multiplied by the conversion factor, i.e. 2 * 1.7 = 3 (rounded).*
6. *The result of step 5, i.e. 3, is added to the lower uniform mark grade boundary, i.e. 60. Therefore, the candidate's total uniform mark is 60 + 3 = 63.*

There are many different UMS which can be constructed ranging from a simple point per grade to the regime currently used for GCE and unitised GCSE specifications. It has been shown that matching uniform marks to raw marks as far as possible reduces the number of anomalies. Too short a scale and approximation and loss of information gives rise to a reduction in percentage at the top end of the grade range and considerable unfairness because the quality of the raw mark is not taken into account. Too long a scale implies spurious discrimination and has proved difficult to explain to centres.

The effect of approximating the raw mark too much is demonstrated in the next example:

In a three unit GCSE, a candidate achieves 74, 89 and 68 marks respectively. In this very simple example, the candidate would get an A grade from the raw mark or from the UMS (using 100 UMS for each unit with boundaries at 90%, 80% and so on) but not using a points conversion. This is because the quality of the grades – one just below the A boundary and the other units just below A – is not recognised.*

Table 1: Raw marks to points conversion

	Unit 1	*Unit 2*	*Unit 3*	*Total*
Raw max	100	100	100	300
Raw A*	80	90	69	239
Raw A	75	81	64	220
Raw B	70	72	59	201
Point max	8	8	8	24
Points A*	8	8	8	24
Points A	7	7	7	21
Points B	6	6	6	18
Candidate raw	74	89	68	231
Candidate points	6	7	7	20
Candidate UMS	78	89	88	255

Uniform grade boundaries in GCSE and GCE specifications

Uniform grade boundaries in GCSE and GCE specifications are established by inter-awarding body agreement. Table 2 shows the mark grade boundaries as percentages of the maximum uniform mark for the unit (or module).

Most GCE subjects are currently based on a 600 uniform mark total. Therefore, the uniform mark grade boundaries for an Advanced

Table 2 Uniform grade boundaries: GCSE and GCE

	GCSE							
Grade	*A**	*A*	*B*	*C*	*D*	*E*	*F*	*G*
%	90	80	70	60	50	40	30	20

	GCE				
Grade	*A*	*B*	*C*	*D*	*E*
%	80	70	60	50	40

specification are A: 480 (= 80% of 600), B: 420 (= 70% of 600), C: 360, D: 300, E: 240. For an evenly balanced scheme of six, equally weighted units, each unit attracts a maximum mark of 100 uniform marks after conversion, with 80 for an A, 70 for B and so on. This gives an A range of 20% of the uniform mark range, with the other pass grades all having the raw grade range mapped on to 10 marks. If the units are not equally weighted or totalling six in number, or both, the UMS for each unit is usually calculated to be in the proportion of that unit of 600, with the boundaries set accordingly, so that in all such cases there will still be greater compensation for an A than any other grade. Table 3 shows this more explicitly for the commonest weightings of 15%, 16.7% and 20%.

Table 3: UMS for GCE specifications

Grade	*Percentage of maximum UMS*	*Specification*	*15% weighted unit*	*16.7% weighted unit*	*20% weighted unit*
Max	100	600	90	100	120
A	80	480	72	80	96
B	70	420	63	70	84
C	60	360	54	60	72
D	50	300	45	50	60
E	40	240	36	40	48

Whatever combination of weighted units are added together (provided the total weighting is 100%), the percentage of marks at each grade boundary will be the same. Therefore, five 20% weighted units will equate, in percentage terms to six 16.7 % weighted units or any other combination. The specification boundary marks will always be the same. In fact, there are almost always six units in a GCE examination, but the weightings are in a variety of combinations.

In September 2008 new GCE courses will start with the first candidates taking A2 examinations in 2010. Most of these will consist of 4 units with a total uniform mark out of 400, although percentages will remain unchanged, that is, 80% of the uniform marks available will determine the A boundary and 40% the E. However, a major challenge will be the introduction of the new A* grade. This is to be awarded to candidates gaining an A grade overall and 90% of the uniform marks available on the A2 units (the second half of the A level). Ensuring fairness and comparability for A* candidates will depend critically on the conversions above A.

Conversions are similar for untiered GCSE assessments, with 90% of the available range assigned to A* with 10% grade bandwidths down to 20% for a G. Maximum uniform marks are not prescribed and are usually chosen as a best fit with the assessment structure. Although the same UMS applies for tiered specifications there are some differences because of the tiers. The maximum uniform mark for a foundation tier unit will be one uniform mark below the B boundary, and the allowed E on the higher tier is set at half the uniform mark grade bandwidth below D (Table 4).

Table 4: Uniform mark boundaries for a 100 ums unit

	Max	*A**	*A*	*B*	*C*	*D*	*E*	*F*	*G*
Untiered	100	90	80	70	60	50	40	30	20
H tier	100	90	80	70	60	50	45		
F tier	69				60	50	40	30	20

Because of the more complex grading regime for GCSE tiered specifications and the additional judgemental boundaries, uniform mark conversions can be more complex, not least because they are potentially different for each unitised GCSE.

It is also important to note that grade boundaries on a uniform mark scale are all the same percentage of the maximum mark. Thus, for GCE assessments, 80% of the maximum mark at both unit and specification level gives the A grade boundary and 40% the E. If it were not so it would be impossible to combine units with different weightings and still maintain the same specification grade boundaries. Table 5 exemplifies the issue.

Table 5: Points conversions for differentially weighted units

Grade	*25%*	*50%*	*75%*	*2@25% +1@50%*	*1@25% +1@75%*
A	6	11	16	23	22
B	5	9	13	19	18
C	4	7	10	15	14
D	3	5	7	11	10
E	2	3	4	7	6
U	1	1	1	3	2

In this very trivial example, and with the points as indicated, the aggregation of differently weighted units leads to different maximum and grade boundary marks. This would be possible to control within a specification which dictated the weighting of each unit, though somewhat confusing; but in a specification, various combinations of units are permitted and such a points regime would be unacceptable.

In all unitised general specifications a grade E is half the value of a grade A at 40% and 80% of the maximum UMS respectively. This relationship is important because a change would affect the basic characteristics of the conversions.

Uniform Mark Scales: challenges and confusions

Uniform marks are not without their difficulties although a range of differing stratagems have been used to overcome the worst. One of the basic issues related to uniform mark use is the maintenance of the value of each raw mark within a unit. No distinction is made, on an assessment's raw mark scale, as to the value of each raw mark and they are, for the purposes of aggregation, all deemed to be of the same value.

The same may not be true after conversion to uniform marks because of the nature of the conversion. If the conversion line is not strictly linear, even if the discontinuities only affect the extreme grades, the consequences are not only undesirable, but also difficult to explain.

In Figure 2 above we see that the conversion line is discontinuous and the conversion rate differs depending on the position of the raw mark relative to the grade boundaries. Grading rules will almost always lead to a line which is segmented, that is, piecewise continuous. For GCE there are two judgemental grades, A and E, and if they are not set exactly at 40% and 80% of the raw mark scale then the line will be discontinuous. Intermediate grades are arithmetic (i.e. determined mathematically), and if the number of marks between A and E is not exactly divisible by 4 (the number of intervening bandwidths) there will also be discontinuities in the line between A and E. For GCSE there are more judgemental boundaries, A, C and F (with D on the higher tier of a tiered unit) and more arithmetic boundaries to be set with a greater likelihood of several conversion factors for the raw marks being applied.

In order to minimise these differences Cambridge Assessment has long had a policy of targeting grade boundaries to align with the UMS scale. So, in writing GCE papers, for example, the question setter will aim for a minimum 'A' performance at around 80% of the raw marks and a minimum 'E' performance at about 40% of the raw marks. The problem is even greater with GCSE assessment because there are more boundaries to be set and potentially different conversion rates between grades which can lead to unpredictable consequences. For tiered specifications with a continuous scale through the tiers it is impossible to set targeted boundaries which will lead to a continuous line, the aim there is to minimise the discontinuities as far as possible.

One of the problems that arose as a result of unequal conversion rates in earlier modular schemes was a reduction in the expected number of high grades. Part of this was due to the lack of consideration of the effect of UMS conversions when mark schemes were devised and often much of the compensating power of a 20 mark UMS A range was lost because raw mark ranges above A were too long, or were not fully utilised. The other reason was the effect of regression. Awarding bodies have addressed the former by the use of a capping mechanism.

Figure 3 shows the effect of capping. Continuing the line showing the conversion from raw to uniform marks between A and E (the dash-dotted line) so that the raw marks above A retain their value, it can be seen that the line reaches the maximum uniform mark before it reaches the maximum raw mark. This is the effect of capping. Without this intervention, candidates gaining raw marks above the A boundary would have conversions based on the dotted line which would not give as much value to their raw marks.

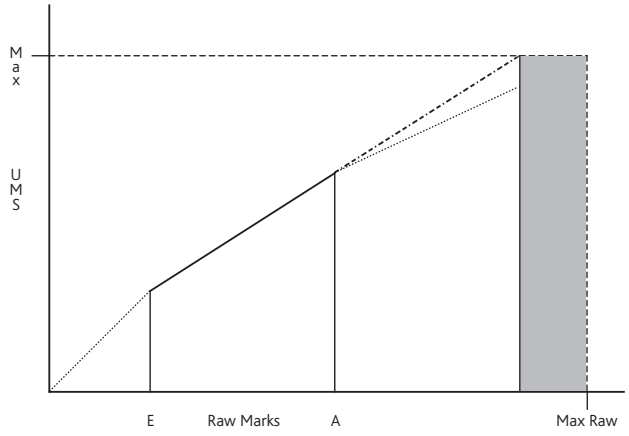


Figure 3: The effect of capping

Candidates in the shaded part of Figure 3, whilst not attaining maximum raw marks, will receive maximum uniform marks. The issues relating to capping can be best illustrated through the following example:

A two tiered GCSE unit has raw marks of 60 in each tier. The maximum uniform mark for the unit is 100. The A, C and D boundaries on the higher tier are 42, 31 and 26 respectively; the foundation tier C and F boundaries are 47 and 28 respectively. The maximum uniform mark for the foundation tier is 69 (one below B). The allowed E will be 45 i.e. 50 – (60–50))/2.

- *On the Higher tier, the B boundary is 36, which makes the A*boundary 42. The tier will be capped at 48 raw marks. All candidates scoring 48 and above on this tier will receive 100 uniform marks. The allowed E is at 23 raw marks and this will be mapped to 45 uniform marks.*
- *On the Foundation tier, the D boundary is 40, so this tier will be capped at 54 and all candidates gaining 54 or more raw marks will receive 69 uniform marks.*

Capping can, however, have undesirable consequences if the full raw mark range has been used. Usually the reason for a low A boundary is that high marks are unattainable so the fact that candidates with marks not on the maximum, but close to the maximum, will be given full uniform marks is not an issue. This might very well happen with the introduction of 'stretch and challenge' questions ('stretch and challenge' questions constitute a potentially promising solution to the issue of high achievement recognition although they are not without their challenges). However, if there is a low A boundary, but the full mark range has been utilised, there may well be significant numbers of candidates on the maximum uniform mark who have achieved very much less than the maximum raw mark.

Capping will also occur when there is a high E boundary even if A is about the 80% mark because twice the A/B distance will be shorter than the maximum raw mark. There is also another issue with a high E boundary. Conversion rates below E will be less than one and the effects of greater value being given to raw marks above E can lead to an unexpected increase in numbers passing the unit. For this reason, for GCE units, a notional N grade has been introduced to ensure that conversion through the E boundary is linear.

A second factor affecting marks at the top and bottom of the grade range is regression which is more accurately known as *attenuation of variance*. This is due to a bunching of marks on aggregation resulting in a reduction in the percentage of candidates gaining the top grades compared with the mean percentage taken over all units. The reverse effect is seen at the bottom of the grade range with an increase in the percentage of E grades. Neither effect is as a result of UMS conversions, although it might be exacerbated as described above.

One of the criticisms which attaches to the UMS method of aggregation is its *invariance*. Specification (and unit) boundaries are pre-defined and thus not open to 'statistical and technical' adjustment post hoc such as may be found with linear schemes. If such variation year-on-year were allowed, then this could give rise to a third type of error. Candidates with the same uniform mark total could be getting different specification grades from year to year. Since raw grade boundaries are set

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

Conclusions

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

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

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

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

References

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

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

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

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

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

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

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

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

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

CAMBRIDGE INTERNATIONAL EXAMINATIONS

The CIE Research Agenda

Stuart Shaw CIE

Introduction

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

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

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

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

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

1. Reliability and validity

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

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

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

2. Comparability and standards

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

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

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

3. New technology

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

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

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

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

4. Development of new CIE products and procedures

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

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

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

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

5. Marking and awarding

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

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

6. Commissioned research

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

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

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

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

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

Research News

Conferences and seminars

House of Commons Research Seminar

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

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

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

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

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

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

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

British Educational Research Association Conference (BERA)

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

European Conference on Educational Research (ECER)

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

International Association for Educational Assessment (IAEA)

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

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

Association for Educational Assessment – Europe (AEA-Europe)

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

Forthcoming conference

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

Publications

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

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

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

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

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

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

Cambridge Assessment
1 Hills Road
Cambridge
CB1 2EU

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

ISSN: 1755–6031