What attracts judges' attention? A comparison of three grading methods

Jackie Greatorex, Nadežda Novakovic and Irenka Suto
Cambridge Assessment

Paper to be presented at the Annual Conference of the International Association for Educational Assessment, Cambridge, September 2008

Contact:
Dr Jackie Greatorex
Core Research Group
Research Division
Cambridge Assessment
1 Hills Road
Cambridge
CB1 2EU
Direct dial. 01223 553805
Fax. 01223 552700
Email: greatorex.j@cambridgeassessment.org.uk
www.cambridgeassessment.org.uk

Abstract

The purpose of grading procedures is to set grade boundaries for various qualifications and to maintain standards that are consistent over time and comparable across units and specifications.

'Thurstone pairs' and 'rank-ordering' are two methods that have been recommended as alternatives to the UK's traditional and principal grading method (Limen referencing) (Pollitt and Elliott, 2003a and b; Kimbell et al, 2007). The traditional method is currently used to grade the national public examinations generally taken by 16- and 18-year-olds. Thurstone pairs and rank-ordering are already well-established as methods in comparability studies in the UK and internationally.

An established method of investigating validity is to explore the appropriateness of features that judges attend to in candidates' work when making grading decisions. Our literature review identified a number of diverse features, which vary in their appropriateness and relevance to the decisions being made. There is no comparable research for Thurstone pairs and rank-ordering.

The aim of this paper is to compare the features of candidates' work that judges attend to when making decisions using the three different grading methods. 'Think aloud' data was collected from judges who used these methods in experimental settings. The implications for comparability studies as well as the alternative grading practices will be discussed.

Introduction

In England, three main awarding bodies (examination boards) offer qualifications in a wide range of subjects to most of the nation's 16- to 19-year olds. As public examinations[1] are 'high stakes', affecting opportunities in the workforce and higher education, questions frequently arise concerning the comparability of examination standards. There is a public expectation that the awarding bodies' examinations are equally difficult and that it is not easier to obtain a particular qualification from one awarding body than from another. Moreover, it is expected that examination standards do not fluctuate over time.

Every year, awarding bodies must determine grade boundaries for their examinations; that is, they must decide the lowest mark on an examination for which candidates will obtain each grade (A, B, C, etc). As examination questions and candidatures change from year to year, this is not easy. Aspects of grading procedures are stipulated by a national regulator (Ofqual, 2008) and awarding bodies work within them, inviting their most senior examiners to participate in decision-making processes. Although several distinct methods for maintaining and monitoring comparability exist (Newton et al, 2007), only one is currently used annually by awarding bodies to maintain year-on-year examination standards. However, other methods have been used in inter-board and international comparability studies (Newton et al, 2007, Arlett, 2003; Greatorex et al, 2003; Edwards and Adams, 2002, 2003; Guthrie, 2003, Townley, 2007). In this paper, we consider three grading methods: 'current awarding', Thurstone pairs, and rank ordering. We explore their validity through investigating potential differences among them relating to the parts of candidates' scripts that receive most attention during decision-making processes.

Current awarding

The UK's traditional and principal grading method is commonly referred to as 'awarding' and utilises limen referencing (Christie and Forrest, 1982; French et al, 1988; Greatorex, 2003). A full description of awarding can be found in the English regulator's Code of Practice (QCA, 2008). Recommendations for grade boundaries are made by a committee of senior examiners (essentially, judges), who meet once the examination marking for a particular syllabus has been completed. Generally, the process is divided into stages:

(1) The committee hears a report from the Principal Examiner about how the question paper performed.
(2) The judges look at examples of candidates' work, starting at the top of the mark range in which the grade boundary is expected to be. They judge the mark at which it is doubtful that the candidates' work is worthy of the grade under consideration. Next, the judges consider candidates' work, starting at the bottom of the mark range, and judge the mark at which it is doubtful that the work is *not* worthy of the grade under consideration. This usually results in a range of marks, known as the 'zone of uncertainty.' (More unusually, a single mark is chosen and recommended as the grade boundary.)
(3) The committee use their "*collective professional judgement*" (Ofqual, 2008), referring to statistics and other information, to recommend an appropriate grade boundary within the zone of uncertainty.

For most public examinations, two or three grade boundaries are determined judgmentally in this way; the remainder are determined arithmetically[2]. Throughout the process, judges have access to 'archive' scripts from the previous year's examination, with marks on the equivalent

[1] Most assessments in qualifications for 16 to 19-year-olds are examinations. However, some assessments are *coursework*, which can include: "extended essays, investigations, practical experiments or performance work" QCA (undated, 2). Unlike examinations, coursework is usually marked by each candidate's own teacher. The coursework marking is then moderated or checked by the awarding bodies.
[2] The procedures for determining grade boundaries for coursework can differ somewhat from the procedures used for examinations.

grade boundary. The marks that scripts received are always visible. Statistical information on the overall performance of the examination and individual questions within it may also be available[3].

Thurstone pairs
In this method, which has been used in comparability studies in the UK and internationally, judges individually compare pairs of candidates' scripts from two different examinations (for example, from two different years, or from two different awarding bodies). For each of many pairs of scripts, the judge must decide which candidate's performance is better (no ties are allowed). The scripts are often cleaned of marks, which are on or near the grade boundary under consideration. Decisions are collated and analysed statistically, enabling all scripts from both examinations to be placed on a single scale of measurement; the standards of the different examinations can then be compared and equivalent marks (and grade boundaries) calculated (Bramley, 2007). For discussion, see Gray (2000), Greatorex et al (2003) and Bramley (2007).

Rank ordering
Like Thurstone pairs, this method has been used in comparability studies, and judges individually compare candidates' scripts (which have been cleaned of marks) from two different examinations. However, rather than judging which of a *pair* of scripts is better, the judge must rank a pack of scripts (e.g. N = 10) in order of overall quality. Half the scripts in the pack are from one examination and the other half are from the other examination. Judges repeat the process with a number of packs of scripts, and scripts from the whole range of marks are used. The script rankings are converted into paired comparison style data. As with Thurstone pairs, statistical analysis enables all scripts from both examinations to be placed on a single scale of measurement; the standards of the different examinations can then be compared and equivalent marks (and grade boundaries) calculated. Whilst rank ordering requires the simulation of some paired comparisons, it is a faster process than Thurstone pairs. For further descriptions of rank ordering and discussions about the method see Bramley (2005), Black and Bramley (in press) and Gill et al (2007).

Research literature on grading methods – what attracts judges' attention?
The judgements made in the above three grading methods differ in nature from method to method. For example, whereas Thurstone pairs and rank ordering call for holistic comparisons of scripts, current awarding demands that judges maintain an internal standard for a particular grade and use it to recommend a grade boundary. More subtle variations across the methods may relate to aspects of the scripts that judges attend to. Although a significant body of research in this area exists on current awarding, an equivalent literature for Thurstone pairs or rank ordering has yet to build up.

In one of the largest studies of current awarding, Murphy et al (1995) found that it was quite common for judges on awarding committees to be invited to look at the answers to key questions. It was suggested that this would give a good guide to the overall achievement of the candidate. Murphy et al argued that this seemed to limit the scrutiny of some judges and that this could lead to a misunderstanding of the totality of the script; looking at only one aspect of a script does not take account of compensations elsewhere in the script. For example, a candidate might do well on a key question but badly on much of the rest of the script, or vice versa. Greatorex (2007) reports on research by Dawes (1979), Einhorn (2000) and Laming (2004) who present similar evidence that "experts are good at knowing what they

[3]The increase in on-screen marking makes it easier for awarding bodies to collect details about performance on individual questions and provide this information to the awarding committee. Additionally, on-screen marking technologies enable scripts to be scanned before they are marked, and some scripts are now presented on screen rather than on paper for consideration by the committee (see *Discussion and Conclusions*). Another recent innovation is that the grade boundary is predicted statistically. The judges are provided with scripts on the recommended boundary and a mark above and below, and are asked whether they would find that boundary acceptable. If not, then the full limen referencing procedure is undertaken.

are looking for but they are not good at mentally combining information" (2007:4). Therefore, focusing judgements on particular questions might be a more successful judgement strategy than holistic evaluations of the whole script, if the questions are a good proxy for the rest of the question paper. After all, such holistic judgements require mentally combining the merits of responses to the variety of questions in a given examination.

In his doctoral dissertation, Cresswell (1997) offered a comprehensive and detailed description of awarding meetings taking place at a British examination board. Cresswell observed and tape-recorded several awarding meetings and noted down the reasons that the judges gave for evaluating the script to be worthy of a certain grade. He divided these reasons into several categories:

- *Objective* reasons - mostly legitimate features to take into account while making grading decisions. These refer to the unity and structure of work ('well organized, 'chaotic'), its complexity ('simplistic', 'rich in contrast'), its content ('something important to say', 'competent review of the issues'), and its intensity ('forcefully argued', 'lively'), with the latter being somewhat problematic in the sense of its appropriateness in evaluating the grade-worthiness of a particular piece of work.
- *Generic* reasons - mostly legitimate reasons for evaluating the grade-worthiness of a candidate's work ('it achieves its aim fully', 'skillfully presented'), although some of these, such as legibility, are more controversial.
- *Affective* reasons – less legitimate features ('gives pleasure', 'interesting')
- *Moral/social* reasons - less legitimate features to give in support of evaluation of student work, especially if they are not linked to features intended to be evaluated.

Crisp (2007) investigated whether assessors pay attention to appropriate features of student work, and whether inappropriate features sometimes influence marking or grading decisions. Crisp asked six examiners to mark 4 - 6 scripts from two Geography exams (an AS unit and an A2 unit[4]) and to complete a grading exercise in which they were asked to set the A/B boundary whilst thinking aloud. The grading exercise simulated 'live' awarding meetings without the potential influence of social or political dynamics. Crisp found that the same types of verbalisations were used both in marking and grading; however, almost all behaviours occurred with much lower frequency in grading than marking most probably because scripts are considered more briefly in grading. Crisp found that most aspects of candidate work noted by examiners were related to geography content knowledge, understanding and skills, which were intended to be assessed. The judges also made frequent reference to Assessment Objectives in the mark scheme, which indicates that they focussed on the appropriate features.

However, Crisp identified a number of more 'problematic features', that is, features one would (arguably) not wish to affect the evaluation of candidates' work. Sometimes, for example, the judges noted the length of response. However, these comments usually related to candidates' responses being shorter or longer than expected, hence not showing sufficient knowledge or including too much information (not directly answering the question). This means that such comments were related to the content of candidates' work after all and were not a sign that judges attended to inappropriate features.

Scharaschkin & Baird (2000) found that degree of consistency of student work influenced grading judgments for biology and sociology A-levels, even though it was not part of the mark scheme guidance. When making grading judgements, judges often say that it is difficult to judge scripts that display inconsistent performance ('rogue' scripts), even though Scharaschkin (1997, cited in Scharaschkin & Baird 2000) showed that consistent performance across papers is highly atypical, and the number of possible mark profiles is enormous. Even

[4] A-levels are taken by numerous 18-year-olds in the UK. Generally, candidates take AS examinations after the first year of study and A2 examinations after the second year of study. (Some assessments are coursework rather than examinations.) The AS and A2 results are combined to give A-level results.

candidates with the same mark profiles may have achieved them in quantitatively different ways. Cresswell (1997) also noted that judges found the scripts that score highly on one question and poorly on another question more difficult to evaluate than the consistent ones.

Scharaschkin and Baird's study involved 24 A level sociology and 17 biology examiners, but not all of these participants were experienced judges. They were asked to make grading judgments, but not to fix the grading boundary on scripts around the grade A (A/B scripts) and grade E (E/N scripts). The researchers divided scripts into three groups according to the consistency of the number of marks that the candidate gained for each question. Therefore, if the candidate gained approximately the same proportion of marks on each question, the script was considered to be consistent. However, if the script gained a high proportion of marks on some questions and not others it was considered to demonstrate inconsistent performance. Scripts that were 'halfway house' were classified as average scripts. Each judge was presented with consistent, average and inconsistent scripts. The judges had access to archive scripts, but no statistical data or PE reports[5] were presented. The results showed that marks and the effects of individual examiners were the strongest contributors to grade judgements, but there was also a significant effect of script consistency. In both mark ranges, sociology examiners preferred consistent performance over inconsistent, and in the A/B range consistent scripts were preferred over average ones. In biology, inconsistent performance was judged to be of a lower standard of achievement than average or consistent performance worth the same number of marks. Overall, the results of the study indicate that the mark consistencies of the random sample of scripts used at an awarding meeting could have a large impact on judges' decisions.

Interestingly, Bramley (2007) suggests that misfitting score profiles (those containing a higher proportion of unexpectedly good answers to difficult questions and/or unexpectedly poor answers to easy questions) could also be considered 'imbalanced', even though they would appear balanced in a Scharaschkin & Baird sense. Such scripts could, in Bramley's view, also pose difficulties for judges. The difficulty of grading inconsistent scripts could be linked to the difficulty of mentally aggregating different levels of performances on different questions. Scharaschkin & Baird mention the psychological work of Shepard (1964) who found that people could not keep two dimensions in mind, and Mynatt et al (1993) who argue that people are poor at holding different states of the world in mind, due to working memory constraints. Greatorex (2007) reports on research by Dawes (1979), Einhorn (2000) and Laming (2004) who present similar evidence that "experts are good at knowing what they are looking for but they are not good at mentally combining information" (2007:4).

As mentioned before, there is not much literature on the script features that judges attend to while using the Thurstone pairs method. However, in their comparability study in A and AS level Geography, Edward & Adams (2002) asked the judges what they paid attention to while making paired comparisons. Concerning the A-level Geography scripts, the judges listed criteria such as depth of understanding (critical evaluation) and sophistication of geographical understanding. Some judges said they took account of the specification when making judgements, with the demand of questions, expectations of mark schemes and the nature of coursework and pre-sight materials influencing their decisions. One judge used intuitive judgement based on experience, while another focused on the synoptic elements of the assessments and how they assessed synthesis, knowledge and understanding. If judges were faced with a tie between two scripts, some judges said they would look into the consistency of the quality of work across all the units, some would focus upon the insightful answers, while some simply went for the gut feeling (Edwards and Adams, 2002). Apart from difficulties arising from comparing scripts from different boards (such as different assessment styles, different levels of demands etc) the judges also reported problems when dealing with imbalanced scripts where candidates have missed out or misread questions, and where there

[5] These are the Principal Examiner's reports about how the examination has worked and are often provided in awarding meetings before the committee make judgements about grade boundaries.

were rubric infringements. Recently, Kimbell *et al* (2007) used a mixture of Thurstone pairs and rank ordering to assess Design and Technology GCSE[6] level work. They used some written criteria to aid the process. However, the practice of using assessment criteria in Thurstone pairs or rank ordering judgements is unusual.

The present study

The present study is the second in a series of linked studies, all investigating three grading methods: current awarding, Thurstone pairs, and rank-ordering. In the first study, Greatorex and Nadas (2008) explored the validity of grading decisions reached whilst 'thinking aloud'. AS-level biology judges were asked to use each of the three grading methods to judge past candidates' scripts from around the grade A/B boundary, both silently and whilst thinking aloud. Analysis of the judgements made in the two conditions indicated that verbalising thoughts made little difference to judges' decisions.

In the present study, we explored the questions (items) that these judges paid most attention to whilst thinking aloud. Our aims were: (i) to identify and compare the numbers and types of questions receiving most attention for each grading method; and (ii) to ascertain how well these questions distinguish between candidates who actually received grades A and B.

Design

The data collection methods have been described in full by Greatorex and Nadas (2008). In summary, the study explored an AS-level biology examination administered by the OCR[7] awarding body: a total of 29 scripts from 2006 and 19 scripts from 2005 were utilised[8]. The participating judges were five senior examiners, all of whom had experience of at least one 'live' awarding meeting for this qualification. In the study, all five judges used all three grading methods, and decisions about past candidates' scripts were made both silently and whilst thinking aloud. Script samples were designed to ensure that each script was encountered no more than once by any particular judge. Overall, however, scripts were used in different total numbers of judgements, as the three grading methods utilised different numbers of scripts.

Each grading method has multiple variants, and in this study, the most established version of each method was used in the research. This meant that for the current awarding method, marks were visible on candidates' scripts, whereas for Thurstone pairs and rank ordering, scripts were cleaned of marks. There was, however, one notable difference between our experimental conditions and operational grading conditions. In current awarding meetings, judges are usually given guidance about which items to consider, whereas this is not the case for Thurstone pairs and rank ordering. We chose not to provide such guidance for any of the three grading methods, as its inevitable influence across the methods would have jeopardised their independence from one another.

Verbal protocol analysis

The verbal protocols generated by the judges whilst thinking aloud were transcribed, then scrutinised for references to items and scripts. References were mostly overt; for example, judges referred to "question 1a" or "3b" or "the question on……" explicitly. However, some transcripts contained some more covert references; for example, judges referred to parts of candidates' scripts that clearly related to a particular item but did not mention the item's number. Other verbalisations could not be linked to any particular item.

Scripts from the two examination years (2005 and 2006) were considered separately. For

[6] General Certificate of Secondary Education (GCSEs) are qualifications obtained by most 16-year-olds at the end of compulsory schooling. They are assessed mostly via examinations (which often require constructed responses), but they also include some coursework.

[7] OCR is one of the UK awarding bodies which is accredited to create and deliver public examinations within the UK. Cambridge Assessment incorporates three awarding bodies, one of which is OCR.

[8] Plus a further 10 scripts from each year, which were used in a warm-up exercise of silent judgement of grading standards.

each grading method and associated script, the presence (either overt or covert) or absence of one or more references to each item was noted. Therefore, items were nested within scripts, and scripts were nested within grading methods. We then calculated the proportion of available scripts on which each item was referenced. This enabled the items to be ranked according to their relative frequencies of referencing (for each grading method).

Identification and comparison of 'most-referenced' items

Table 1 indicates the rankings of questions from the 2006 examination. It can be seen that although the rankings vary slightly from method to method they are reasonably similar and, question 8 was the most referenced question for all three methods. This question was the only one in the paper requiring candidates to give a long answer, and was also the question with the greatest maximum mark. Question 5, which was worth 3 marks and required candidates to give a brief explanation, was the second most referenced question for both current awarding and Thurstone pairs, and the fourth most referenced question for rank ordering.

The bottom of Table 1 reveals question 4 to be the least referenced question for both Thurstone pairs and rank ordering, and the second least referenced question for current awarding. Question 13 was the least referenced question for current awarding, and the second least referenced question for rank ordering. Both questions 4 and 13 were short 2-mark questions requiring candidates to label a diagram.

Table 1: Rankings of questions from the 2006 examination, in order of their relative frequencies of referencing

Current awarding			Thurstone pairs			Rank ordering		
Rank (most referenced to least referenced)	Qu. (Max. mark)	Qu. type	Rank (most referenced to least referenced)	Qu. (Max. mark)	Qu. type	Rank (most referenced to least referenced)	Qu. (Max. mark)	Qu. type
1	8 (6)	Long answer	*1*	8 (6)	Long answer	*1*	8 (6)	Long answer
2	5 (3)	Explain	*2*	5 (3)	Explain	*2*	6 (2)	Explain
3	11 (4)	Explain	*2*	11 (4)	Explain	*3*	7 (2)	Calculation
4	16 (3)	Explain	*2*	12 (3)	Explain	*4*	5 (3)	Explain
5	9 (5)	Gap filling	*2*	16 (3)	Explain	*5*	17 (2)	Explain
5	6 (2)	Explain	*3*	9 (5)	Gap filling	*6*	3 (1)	Multiple choice
6	3 (1)	Multiple choice	*3*	3 (1)	Multiple choice	*6*	11 (4)	Explain
7	1 (1)	Labelling	*3*	6 (2)	Explain	*7*	1 (1)	Labelling
8	2 (1)	Multiple choice	*3*	7 (2)	Calculation	*7*	2 (1)	Multiple choice
8	7 (2)	Calculation	*4*	2 (1)	Multiple choice	*7*	14 (2)	Explain
9	17 (2)	Explain	*5*	1 (1)	Labelling	*7*	16 (3)	Explain
10	12 (3)	Explain	*6*	13 (2)	Labelling	*8*	12 (3)	Explain
11	10 (4)	Gap filling	*7*	14 (2)	Explain	*9*	9 (5)	Gap filling
12	15 (1)	Explain	*7*	15 (1)	Explain	*10*	15 (1)	Explain
13	14 (2)	Explain	*8*	10 (4)	Gap filling	*11*	10 (4)	Gap filling
14	4 (2)	Labelling	*9*	17 (2)	Explain	*12*	13 (2)	Labelling
15	13 (2)	Labelling	*10*	4 (2)	Labelling	*13*	4 (2)	Labelling

How well do the 'most referenced' questions distinguish between candidates who actually received grades A and B?

The item analysis used marks given during 'live' marking. Scripts used in the silent and the think aloud tasks were included in the analysis. For the 2006 scripts used in the study (N = 29), we identified items that distinguish between the achievement of grade A and grade B

candidates. Mann Whitney U tests were used to compare the rank of item marks from grade B candidates with the rank of item marks of grade A candidates (see Table 2). Just two questions (5 and 13) were found to statistically distinguish between grade A and grade B candidates, further vindicating judges' use of question 5 in all three grading methods. However question 13 was one of the least referenced questions; it appears that judges were unaware of its discriminatory power. Surprisingly, question 8 was found not to distinguish between grade A and grade B candidates, despite it being the most frequently referenced question for all three methods.

Table 2: Mann Whitney U tests to compare the rank of item marks from grade B candidates with the rank of item marks of grade A candidates for live (2006) scripts

Question	U	Significance (exact)	Significance level	Maximum mark
1	127.5	0.131	> 0.05	1
2	147.0	0.352	> 0.05	1
3	142.5	0.283	> 0.05	1
4	172.5	0.831	> 0.05	2
5	**88.5**	**0.007**	**< 0.01**	**3**
6	160.5	0.578	> 0.05	2
7	141.5	0.270	> 0.05	2
8	132.0	0.172	> 0.05	6
9	154.5	0.466	> 0.05	5
10	172.0	0.831	> 0.05	4
11	160.0	0.578	> 0.05	4
12	118.0	0.076	> 0.05	3
13	**106.5**	**0.033**	**< 0.05**	**2**
14	122.0	0.097	> 0.05	2
15	144.0	0.309	> 0.05	1
16	166.5	0.700	> 0.05	3
17	157.5	0.521	> 0.05	2

Discussion and conclusions

An evidence-based response to our research question of what attracts judges' attention in grading methods may be: "some key questions but not necessary the most useful ones". Our study of three different methods of grading AS level biology scripts (current awarding, Thurstone pairs, and rank ordering) has revealed that the judges in this research are only partially aware of which questions it is most useful to focus on. Whilst they justifiably utilised question 5 in their judgements, the most referenced question (8) did not statistically discriminate well between grade A and grade B candidates. Conversely, question 13 discriminated well but was referenced by judges relatively infrequently. The rankings of relative frequencies with which questions were referenced were broadly similar across the three grading methods, suggesting that the questions that judges focussed on did not vary substantially from method to method.

The research has several limitations. These include its use of only a small group of judges and small script samples for a single examination subject, any of which may compromise its generalisability. The item (question) level data analysis was conducted on the 38 scripts used in the research, rather than on the performances of the whole candidature. Furthermore, the verbal protocols generated through 'thinking aloud' are unlikely to be complete records of judges' thoughts. However, whilst theoretically, some references to some questions could have been harder to verbalise than others, this seems unlikely. Some questions are undoubtedly harder to *answer* or *mark* than others, but there is no salient reason why verbalisations of shifts of attention to them should vary in difficulty. Concerns relating to the study's ecological validity (the research method was novel to judges) are hopefully allayed by

Greatorex and Nadas's (2008) earlier study of the same individuals and scripts, which indicated that having to verbalise thoughts made little difference to judges' grading decisions.

Nevertheless, the study's findings may have some important implications. First, if it is considered appropriate that judges should focus on the questions that best discriminate, then guidance on which questions to focus on might facilitate the judgement process. Until recently, the use of traditional script-by-script paper-based marking has meant that item-level data has been difficult to provide operationally. However, e-marking has greatly facilitated its collection and analysis, and the provision of statistics about each item is now feasible. Indeed, in recent years, for e-marked examinations, Chief Examiners at the OCR awarding body have been using this information to support other judges in current awarding meetings. One argument for using key discriminating questions is that it avoids judges having to mentally combine information from a variety of questions in the one examination (which they are unlikely to be good at). There are, however, arguments against focussing on particular questions when determining grade boundaries. For example, if the consistency of performance across all questions is a trait that distinguishes scripts of different grades, then judgements must by definition be more holistic. At the root of this issue lie the definitions of performances at different grades, and the distinctions among them that are implicit in a qualification's grade descriptors. Another objection to referring only to certain items is that arguably these items are doubly credited: once during marking and once in awarding.

Although at present, the rank ordering and Thurstone pairs grading methods are not used operationally to determine grade boundaries in England's major public examinations, they have been scrutinised in several recent research projects and their suitability is being considered by assessment professionals. Although both methods request judges to make holistic judgements about the relative qualities of scripts, it is clear from the uneven referencing of questions in our verbal protocol data that not all questions in a script receive equal consideration. Arguably, the provision of statistics about each item would provide guidance on this subset and benefit the judgement process. However, this is clearly an issue for further investigation, since it raises questions surrounding the purpose of the other questions in the examination. Our finding that the questions that judges focussed on did not vary substantially across the three grading methods would suggest that there are no grounds for favouring one method over the others on the basis of the questions focussed upon.

Research about expertise beyond the context of educational assessment demonstrates that experts are good at knowing what to look for, but they are not good at mentally combining information (Dawes, 1979; Einhorn, 2000; Laming, 2004). If we apply these arguments to our context, it seems that senior examiners may know what qualities to heed in candidates' work, but may be less good at integrating information, for example at combining cues from different questions or combining characteristics of the performance such as the consistency of achievement along with the quality of communication. There is research in progress at Cambridge Assessment exploring the extent to which different features of scripts contribute to decisions about grading standards.

Bibliography

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.*

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

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

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

Christie, T. and Forrest, G. M. (1982) *Defining Examination Standards*, Schools Council Research Studies: London.

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

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, Azerbijan, September.

Dawes, R. (1979) The Robust Beauty of Improper Linear Models in Decision Making, *American Psychologist, 34,* 7, 571-582.

Edwards, E. and 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. and 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.*

Einhorn, H. J. (2000) Expert judgement: Some necessary conditions and an example. In: T. Connelly, H. R. Arkes, and K. R. Hammond (Eds). *Judgement and decision making: an interdisciplinary reader* (2nd ed., pp. 324-335) Cambridge University Press: Cambridge.

French, S., Slater, J. B., Vassiloglou, M., and Willmott, A. S. (1988) *The role of Descriptive and Normative Techniques in Examination Assessment*. In: H. D. Black and B. Dockrell (Eds). Monograph of Evaluation and Assessment Series No. 3, Scottish Academic Press: Edinburgh.

Gill,T., Bramley, T. & Black, B. (2007). *An investigation of standard maintaining in GCSE English using a rank-ordering method.* Paper presented at the British Educational Research Association annual conference, Institute of Education, London.

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, Heriot-Watt University, Edinburgh.

Greatorex, J., Hamnett, L. and 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. (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 the British Educational Research Association Conference, 2007.

Greatorex, J. and Nadas, R. (2008) *Using 'thinking aloud' to investigate judgements about A-level standards: does verbalising thoughts result in different decisions?* A paper to be presented at the British Educational Research Association Annual Conference, September 2008, Edinburgh.

Gray, E. (2000) *A comparability study in GCSE science 1998. A study based on the 1998 summer examination.* Organised by Oxford, Cambridge and RSA Examinations (Midland Examining Group) on behalf of the joint forum for GCSE and GCE.

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.*

Kimbell, R., Wheeler, A., Miller, S. and Pollitt, A. (2007), *E-scape portfolio assessment phase 2 report*. Department of Design, Goldsmiths, University of London.

Laming, D. (2004) *Human judgement The Eye of the Beholder*, Cambridge University Press: Cambridge.

Murphy, R., Burke, P., Cotton, T., Hancock, J., Partington, J., Robinson, C., Tolley, H., Wilmut, J. and 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.

Mynatt, C.R., Doherty, M.E. and Dragan, W. (1993). Information relevance, working memory, and the consideration of alternatives, *Quarterly Journal of Experimental Psychology: Human Experimental Psychology,* 46A, 759-778.

Newton, P., Baird, J., Goldstein, H, Patrick, H. and Tymms, P. (2007) *Techniques for monitoring the comparability of examination standards.* Qualifications and Curriculum Authority: London.

Ofqual (2008) [online.] Available at: http://www.ofqual.gov.uk/

Pollitt, A. and 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. and 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.

Qualifications and Curriculum Authority (undated) Coursework a guide for parents [online.] Available at http://ofqual.gov.uk/files/qca-06-3403-csewk-parents.pdf

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

Shepard, R. N. (1964) On subjectivity optimal selection among multi-attribute alternatives, in: M W Shelley & G L Bryan (eds) *Human Judgements and Optimality*, Wiley: New York.

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.