CAMBRIDGE ASSESSMENT

Grading examinations using expert judgements from a diverse pool of judges

Nicholas Raikes[1], Sara Scorey[2] and Hannah Shiell[1]

[1]Research Division	[2]Operational Research Team
Cambridge Assessment	OCR
1, Regent Street	1 Hills Road
Cambridge	Cambridge
CB2 1GG	CB1 2EU
United Kingdom	United Kingdom

Raikes.N@cambridgeassessment.org.uk

Scorey.S@cambridgeassessment.org.uk

Shiell.H@cambridgeassessment.org.uk

A paper presented to the 34th annual conference of the International Association for Educational Assessment, Cambridge, UK, September 2008

Abstract

In normal procedures for grading GCE Advanced level and GCSE examinations, an Awarding Committee of senior examiners recommends grade boundary marks based on their judgement of the quality of scripts, informed by technical and statistical evidence. The aim of our research was to investigate whether an adapted Thurstone Pairs methodology (see Bramley and Black, 2008; Bramley, Gill and Black, 2008) could enable a more diverse range of judges to take part. The key advantage of the Thurstone method for our purposes is that it enables two examinations to be equated via judges making direct comparisons of scripts from both examinations, and does not depend on the judges' internal conceptions of the standard required for any grade.

A General Certificate of Education (GCE) Advanced Subsidiary (AS) unit in biology provided the context for the study reported here. The June 2007 and January 2008 examinations from this unit were equated using paired comparison data from the following four groups of judges: members of the existing Awarding Committee; other examiners that had marked the scripts operationally; teachers that had taught candidates for the examinations but not marked them; and university lecturers that teach biology to first year undergraduates.

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 from the equated June 2007 boundaries, 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 concluded that the examiners, teachers, lecturers and members of the current Awarding Committee made very similar judgments, and members of all four groups could take part in a paired comparison exercise for setting grade boundaries without compromising reliability.

Introduction

Maintaining grading standards

When more than one form or version of an examination exists, a way must be found of equating raw scores so that the outcomes reported are comparable regardless of the particular examination taken.

In the present paper we report the results of an experiment that investigated one way of equating raw scores based on judgements made by a range of content-experts of the relative quality of sample candidate work. The method used required no items or candidates to be common to the examinations equated. The main focus of the research was whether the experts' professional occupations affected the outcomes of their judgements.

Context

Our context was General Certificate of Education Advanced Subsidiary (GCE AS) examinations, generally taken by pupils in England, Wales and Northern Ireland at age 17+ during, or at the end of, their first year of post compulsory education. The research is relevant to similar examinations, however. Key features of the examinations of relevance to the present study are:

- The examinations are high stakes for candidates, teachers and universities, since ultimately candidates' university places may depend upon their results;
- The examinations are content-based and mainly contain constructed response questions ranging from short answers to extended writing, depending on the subject;
- The examinations are generally available for candidates to take on one or two occasions per year;
- Entirely original question papers are used on each occasion – questions are used once only;
- Question papers must pass a rigorous quality assurance process, but no formal pre-testing with candidates occurs;
- Candidates' results are reported as grades, with passing grades from A (top) to E.

How grade boundary marks are currently set operationally

Since April 2008, GCE and other public examinations in England have been regulated by Ofqual, the Office of the Qualifications and Examinations Regulator. Ofqual inherited this responsibility from QCA, the Qualifications and Curriculum Authority. QCA / Ofqual's mandatory code of practice (Ofqual, 2008) specifies the Awarding process by which grade boundary marks are set. The code of practice states (p.33) that the "prime objectives [of Awarding] are the maintenance of grade standards over time and across different specifications within a qualification type." Grade boundary marks are recommended by an Awarding Committee of senior examiners. On page 36, the code of practice states that "Each boundary must be set using professional judgement. The judgement must reflect the quality of candidates' work, informed by the relevant technical and statistical evidence." Sample scripts are inspected by the Awarding Committee when determining the A and E lower boundaries (termed "key boundaries"). The code of practice specifies the following procedure (p. 37):

> Awarders must first consider candidates' work in the range for each key boundary, ensuring that a sufficient amount of candidates' work is inspected. They must consider each mark in turn, as follows.
>
> i First, working down from the top of the range, awarders must identify the lowest mark for which there is consensus that the quality of work is worthy of the higher grade of the boundary pair. This forms the upper limiting mark.
>
> ii Next, working up from the bottom of the range, awarders must identify the highest mark for which there is consensus that the quality of work is not worthy of the higher grade. The mark above this forms the lower limiting mark.
>
> Awarders must then use their collective professional judgement to recommend a single mark for the grade boundary, which normally will lie within the range including the two limiting marks. ... All awarders must then consider candidates' work at the recommended mark to confirm that this is appropriate and to identify scripts to be archived.

The procedure therefore depends on Awarders making an absolute judgement of the quality of scripts, e.g. "this script is clearly worth an 'A'", "this script is clearly **not** worth an 'A'", etc. Each Awarder relies on an internal, abstract standard developed over time based on experience and prior inspection of archive scripts, etc.

The B, C and D grade boundaries are set arithmetically by dividing the raw mark range between the A and E boundaries into four.

Paired comparison methods for standard maintaining

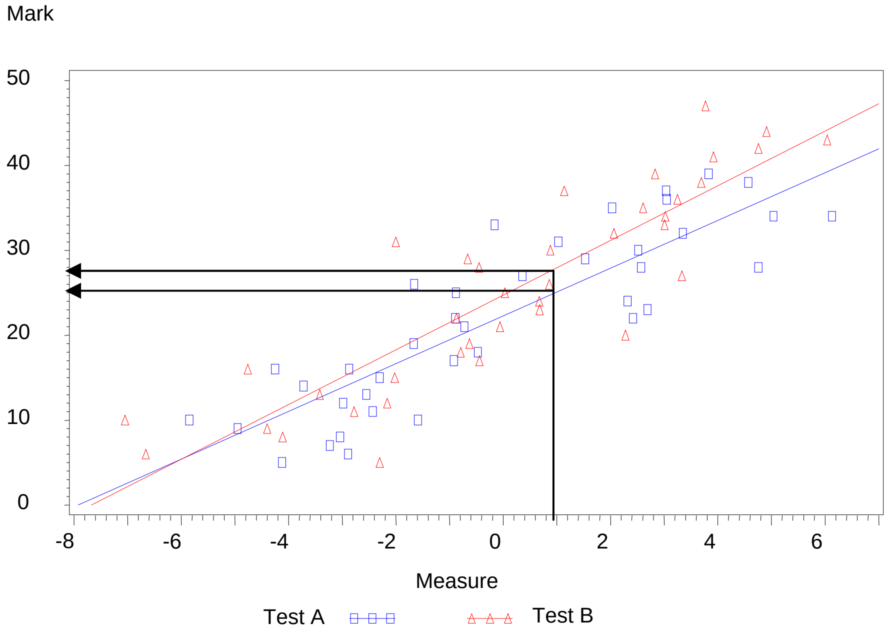
Thurstone (1927a, 1927b) introduced methods for constructing an interval scale and simultaneously locating objects on the scale using a process of pairwise comparisons by judges.

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* judgments about the objects in a pair, even if their absolute judgments 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 – i.e. 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.

Application of Thurstone's work to equating examinations involves constructing a single interval scale for the examinations and simultaneously placing sample scripts from these examinations onto this scale. Since the scripts all have raw marks, lines of best fit can be drawn for each examination that link raw mark with scale measure. Figure 1, taken from Bramley (ibid), illustrates this for two tests, A and B: each square represents a sample script from Test A, each triangle a sample script form Test B, and the best fit lines relate scale measure to raw mark. The actual scale values are arbitrary, but they enable raw marks from the different tests to be equated. If one of the tests – Test A, let's say – was a previously administered test with grade boundary marks already established, and the other test (Test B) was a new test, comparable grade boundary marks for Test B can be obtained by reading off the Test B marks that correspond to the Test A grade boundaries. Note that the method equates the **entire range of marks**; every grade boundary mark can be equated without the need to interpolate between A and E, as in the current operational procedure.

Figure 1: Example equating of two tests (from Bramley, 2007)



Bramley (ibid) explains that the chart presented here as Figure 1 was actually based on a study involving an adapted Thurstone Pairs method, where judges were presented with more than two scripts at a time and asked to rank them. A practical drawback of using pairs of examination scripts is that scripts are typically eight or more pages long and take a considerable time to read; judges can be become very bored seeing the same scripts multiple times in different pairing combinations. This drawback led Bramley (2005) to use an adaptation of the Thurstone Pairs method where ten scripts were presented at a time, and the judges instructed to rank the scripts in order of their perceived quality. Bramley (ibid) recommends treating the rankings as though they came from paired comparisons (i.e. 1st beats 2nd, 1st beats 3rd, etc.), and presents some evidence that the lack of local independence of the inferred paired comparisons has little practical effect on the measures, though the standard errors of the measures are muted (i.e. the scale appears a little more reliable than it actually is). Bramley, Gill and Black (2008) provide further evidence of the validity of this “rank-ordering” method for standard maintaining.

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. Collect feedback from participants about how difficult they found the task, how long it took them and their confidence in their decisions
4. 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. The present paper reports results for aims 1 and 2 for biology only. Work continues on sociology and the other aims.

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 singe 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 decided to use real scripts from the live examinations in the range 14 – 52 raw marks. This extended 6 marks below the lower E boundary and 6 marks above the A boundary set operationally in June 2007. This range comfortably encompassed all E and A boundary marks set operationally for the last six examinations, so we were confident that the January 2008 boundaries would lie within this range.

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

Members of the Awarding Committee and examiners were recruited via a personal email. Teachers were recruited via letters to Heads of Biology at centres that entered candidates for the assessment. Lecturers were recruited via emails to Heads of Biology Departments at universities. The following numbers of participants completed the exercise and returned materials:

[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.

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, i.e. 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, i.e. 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.

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 X 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 X 3 script-copies divided by 78 scripts = 15.4 triples per script);
- 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 – i.e. 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

[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

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 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.

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

	Judges n	Triples n	Pairs n	Reliability*	Correlation between raw mark & measure		Paired comparisons consistent with measures
					Jun	Jan	
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

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

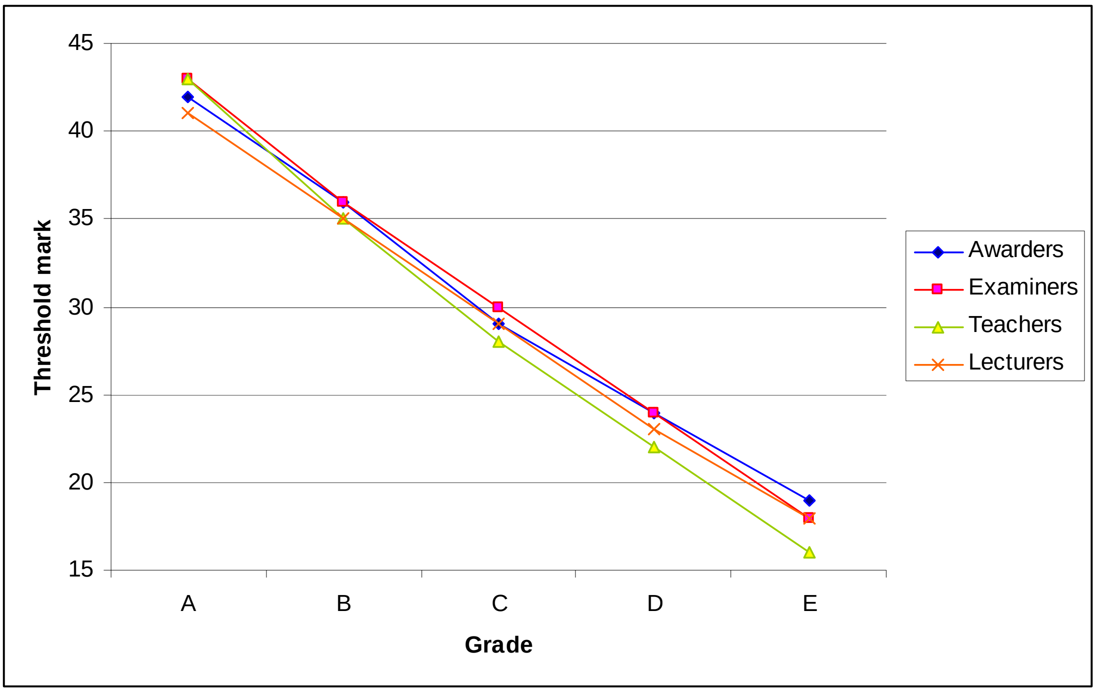
Table 3 gives the grade boundary marks estimated from each group's rankings for the January 2008 examination. Figure 2 presents the same information graphically (the lines between the points have been drawn in for clarity but have no meaning). The figures are similar for each group, with a maximum spread 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 2: 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. 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 judgments. 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, Beth (2008). *Using an adapted rank-ordering method to investigate January versus June awarding standards*. Paper presented at the Fourth Biennial EARLI / Nortumbria Assessment Conference, Berlin, Germany, August 2008.

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

Bramley, Tom (2007). *Paired comparison methods* Chapter 7 of 'Techniques for monitoring the comparability of examination standards' edited by Newton, Paul; Baird, Jo-Anne; Goldstein, Harvey; Patrick, Helen and Tymms, Peter. Published by the QCA, London, UK.

Bramley, Tom and Black, Beth (2008). *Maintaining performance standards: aligning raw score scales on different tests via a latent trait created by rank-ordering examinees' work*. Paper presented at the Third International Rasch Measurement conference, University of Western Australia, Perth, January 2008.

Bramley, Tom; Gill, Tim and Black, Beth (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, Donald (2004). *Human judgment: the eye of the beholder*. London: Thomson

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

Ofqual (2008). *GCSE, GCE and AEA Code of Practice*, http://www.ofqual.gov.uk/files/Code_of_practice_April_2008.pdf Accessed 11 July 2008.

Thurstone, L.L. (1927a). *Psychophysical analysis*. American Journal of Psychology, 38, 368-389. Chapter 2 in Thurstone, L.L. (1959). The measurement of values. University of Chicago Press, Chicago, Illinois.

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

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