# 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 |

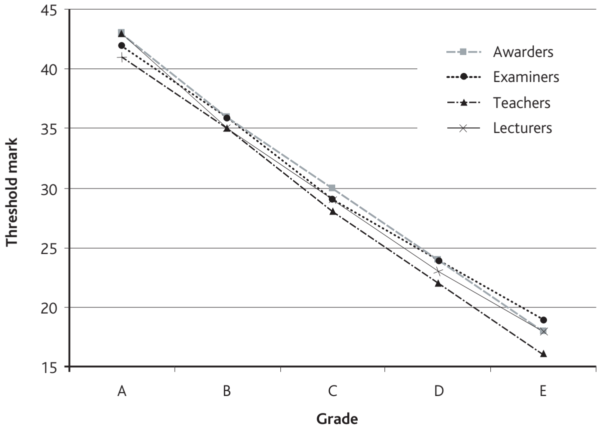
### Estimated grade boundaries for January 2008

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.

**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**



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.

## 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.

ASSESSMENT JUDGEMENTS

# 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.