The reliabilities of three potential methods of capturing expert judgement in determining grade boundaries

Nadežda Novaković and Irenka Suto Research Division

Introduction

In England there is a strong public expectation that qualification standards should remain constant over time. For example, a candidate who achieves a grade B in GCSE Spanish in one year should be considered 'comparable' in some sense to candidates from previous years who also achieved a grade B in GCSE Spanish. At each examination session, awarding bodies must therefore determine the grade boundaries for their examinations that equate to those of previous sessions. A great deal of research activity is directed towards investigating different methods for capturing the expert judgement of professionals who are given the responsibility of determining grade boundaries and thus maintaining year-on-year examination standards.

In this article, we report the results of some research[1] investigating the reliabilities of three such (potential) methods for capturing expert judgement, as used in:

(i) Traditional (current) awarding

(ii) Thurstone pairs

(iii) Rank ordering.

The traditional awarding method is the principal method used operationally for determining grade boundaries in the context of public examinations and England, while rank ordering and Thurstone pairs have been sometimes suggested as alternatives to the judgemental process used in traditional awarding.

Traditional awarding

When the traditional awarding method is used, a committee of senior examiners (led by a Chair of Examiners) looks at a sample of candidates' scripts in the mark range where the grade boundary is expected to be. They are required to make holistic, absolute judgements about whether each script on a particular mark is worthy of the grade in question, for example, 'this script is worthy of grade A' or 'this script is a borderline grade B script'. This type of judgement implies that examiners (judges) have an internal standard about what, for instance, a grade A script should look like; it is assumed that judges would have internalised this standard partly through experience and partly through studying archive scripts.

The judges decide on the lowest mark for which there is consensus that the work is worthy of the higher grade and the highest mark for which there is consensus that the work is not worthy of the higher grade. This gives a range of marks called the 'zone of uncertainty', or simply 'zone'. The committee then use their collective professional judgement, referring to statistical information on the overall performance of the examination, to recommend an appropriate grade boundary from within that range. Throughout the process, judges have access to 'archive' scripts from the previous year's examination, with marks on the equivalent grade boundary. Statistical information on performance on individual questions may also be available.

Concerns have been raised over the reliability of the judgements made in the traditional awarding method (Willmott & Nuttall, 1975; Greatorex & Nádas, 2008). Good and Cresswell (1988) replicated some awarding meetings for GCSE French, History and Physics and found that parallel groups of judges reached slightly different decisions about grade boundaries, which, if substituted for one another, would have affected the grade of 13% of French candidates, 17% of physics candidates and 38% of history candidates. Imperfect reliability may stem from the method's reliance on absolute judgements. Drawing on Laming's theory of the nature of human judgement (2004), that absolute judgement cannot occur and that all judgements are comparisons of one thing with another, Raikes *et al*. (2008) have argued for replacing traditional awarding with methods in which judges make relative judgements about the quality of candidates' work.

A recent empirical study by Gill and Bramley (2008) supports this view. The study's participants were experienced history and physics examiners who were given pairs of scripts and asked to make absolute judgements about the grade each script deserved. The participants also made relative judgements about the pairs of scripts, that is, they judged which of the two scripts was better in terms of overall quality. All scripts were cleaned of marks and the participants had no reference to archive scripts or any statistical information. The examiners' judgements were compared with the marks and grades that the scripts originally received, and the results showed that examiners had difficulty in replicating the decisions made at the live awarding meetings which they themselves had attended: the percentage of judgements matching the original grades was below 40% for history and below 25% for physics. On the other hand, the overall accuracy of the relative judgements was higher than that of the absolute judgements (history examiners ordered 66% of the paired comparisons in correct mark order, while physics examiners ordered 78% of the comparisons in correct mark order).

Methods using relative judgements

In view of the criticisms levelled against the traditional awarding method, Thurstone pairs and rank ordering have been suggested as possible replacement methods of capturing expert judgement in determining

1 The wider research project also addressed an aspect of the validities of these methods by investigating and comparing the features of candidates' work that most influence experts in each method; these results were presented by Novaković and Suto (2009).

grade boundaries. Both methods rely on examiners making relative holistic judgements about the quality of candidates' work, which arguably have more psychological validity than absolute judgements. Furthermore, judgements made in rank ordering and Thurstone pairs are not influenced by statistics or by candidates' marks, which are always visible in traditional awarding (see Black & Bramley, 2008, and Greatorex, 2007 for a detailed list of advantages of rank ordering and Thurstone pairs over the traditional awarding method).

Thurstone pairs

In recent decades, the Thurstone pairs method (Thurstone, 1927a, b) has been used in comparability studies in the UK and internationally. In this method, judges are required to individually compare pairs of candidates' scripts from two different examinations (for example, from two different years). 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. If these comparisons are repeated many times, then Rasch analysis can be used to place all scripts from both examinations on a single common scale of measurement, representing a latent construct of script quality. The equivalent marks of the different examinations can then be calculated, enabling standards to be compared (see Bramley, 2007).

Kimbell *et al*. (2007) are the first to have investigated the use of Thurstone pairs as a method for harnessing expert judgement in grading, but no systematic comparisons with the outcomes of more conventional methods of grading have been carried out. Hence, there are no established procedures for using Thurstone pairs in grading. The main drawback of the Thurstone pairs method is that it can be time consuming, particularly when considering a large number of scripts, which take time to read and might be remembered, thus probably violating the requirement that each paired comparison should be independent of any previous comparison.

Rank-ordering

The rank ordering method (Bramley, 2005) is similar to Thurstone pairs in that 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 individual scripts in a pack, 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. Each judge has a different combination of scripts in their packs. As with Thurstone pairs, Rasch analysis enables all scripts from both examinations to be placed on a single scale of measurement; the equivalent marks (and grade boundaries) can then be calculated. Rank ordering is more time-efficient than Thurstone pairs and it can be designed to ensure that the number of times a judge sees a particular script is minimised, reducing the possibility of the scripts being remembered.

The rank-ordering method has been used for the purposes of setting grade boundaries, both in an operational setting (for Key Stage 3 English examination, see Bramley, 2005) and in research settings (Black, 2008; Black & Bramley, 2008; Elliott *et al*., 2005; Gill & Black, 2006).

Black and Bramley (2008), and Gill *et al*. (2007) have investigated whether traditional awarding and rank ordering generate the same grade boundaries, by using these two methods to cross-validate the traditional awarding of an A-level psychology paper and GCSE English paper respectively. Both studies found some concurrence and some disparity at key grade boundaries. However, given that traditional awarding uses a blend of both judgemental and statistical information, the methods' outcomes should not be expected to be identical.

An adaptation of the rank ordering method has recently been used experimentally by Raikes *et al*. (2008) in the context of an AS-level biology examination. Research participants were required to judge the relative qualities of sets of three scripts at a time. Four groups of judges were involved in the study: members of the existing awarding committee; other examiners who had marked the scripts operationally; teachers who had taught candidates for the examinations but not marked them; and university lecturers who teach biology to first year undergraduates. Raikes *et al*. identified very high levels of intra-group and inter-group reliability for the scales and measures estimated from all four groups' judgements.

The present study

We conducted a three-way comparison of the intra-method and inter-method reliabilities of all three methods in the context of setting grade boundaries.

Intra-method reliability refers to the comparison of the grade boundaries yielded by each single method in turn, if used by different groups of judges and on different sets of scripts. While the literature indicates that the intra-method reliability of traditional awarding is imperfect, it is unclear how it compares with that of the Thurstone pairs and rank ordering methods when these are used in grading. To our knowledge, a direct comparison has not previously been made.

Inter-method reliability refers to the comparison of the grade boundaries that the three methods would yield if used on the same examination papers. Arguably, high inter-method reliability would suggest that judgements are made in reference to a common construct (or a common subset of constructs). The above-mentioned studies by Black and Bramley (2008), and Gill *et al*. (2007) have addressed this issue to some extent by comparing the outcomes of the traditional and rank ordering methods. However, this issue is clearly ripe for further investigation. The Thurstone pairs method has not been compared directly with either of the other two methods in the context of standard maintenance.

Experimental design

The research focused on two written examination papers with contrasting question and response styles, which were administered by OCR examinations in June 2007 (available from www.ocr.org.uk). One paper (maximum mark = 45) was from an AS-level biology syllabus, and the other paper (maximum mark = 90) was from a GCSE English syllabus. The research was carried out using samples of past candidates' scripts: for biology, the research focussed on the E/U and A/B grade boundaries; for English, the research focussed on the C/D and A/B boundaries.

The experimental design was identical for biology and English, taking the form of a 3×3 'Latin square' (see Table 1). For each subject, three mutually exclusive sets of examination scripts were created, which were matched for mark. Three groups of ten 'judges' (examiners, matched for experience of the methods) made judgements using each of the three methods on a different set of scripts. Thus, each judge group encountered the three methods in a unique order, and ultimately, judgements of each method were conducted on all three script sets. The Latin square design

thereby enabled comprehensive comparisons of the three methods, whilst controlling for order effects.

Table 1: Latin square design

Judge group	*Script set and order of attempting tasks*		
	1	*2*	*3*
1	Rank ordering	Traditional awarding	Thurstone pairs
2	Thurstone pairs	Rank ordering	Traditional awarding
3	Traditional awarding	Thurstone pairs	Rank ordering

Procedure

Each judge received three sets of photocopied scripts (one for each of the tasks) together with a covering letter, detailed instruction sheets for individual tasks, statistical information on the candidates for use in the traditional awarding task, charts for recording judgements, and copies of the question papers and mark schemes from June 2007 and June 2006.[2] The judges were given three weeks to complete the tasks from home and were advised to take about half a day per task. They were asked to (re)familiarise themselves with the question papers and the mark schemes before embarking on the tasks. Judges were asked not to re-mark the scripts; instead, they should make a holistic judgement about each script's quality.

For each task, each group of judges used scripts drawn from a different script set (see Table 1). Within each script set, the numbers and marks of scripts selected for use in each judgemental method were determined by the common practice for that method. (Script selection for Thurstone pairs followed previous studies (Bell *et al*.,1998; Bramley *et al*.,1998)).

Traditional awarding

Biology judges received ten scripts around the E/U boundary and ten scripts around the A/B grade boundary. They also received four 'archive' scripts from June 2006 – two on each grade boundary mark. English judges received twelve scripts around the C/D boundary and twelve around the A/B boundary, as well as four 'archive' scripts – two on each grade boundary mark. The judges' task was to decide whether the June 2007 scripts were worthy of the grade under consideration. The scripts' marks were clearly visible.

Thurstone pairs

For each subject, the judges received two packs of scripts. Pack 1 contained a total of 20 scripts around the higher boundary, while Pack 2 contained a total of 20 scripts around the lower boundary. In each pack, 10 scripts were from June 2006 and 10 scripts from June 2007. The judges compared two scripts at a time, and judged which represented the better performance.

Rank ordering

For each subject, the judges received four packs of scripts. Each pack comprised 10 scripts: 5 from 2007 and 5 from 2006. Each pack contained a unique selection of scripts, but there were common scripts between the judges' packs allowing each entire set of scripts to be linked. The task included all the scripts that were used in the study and these covered the entire mark range for both examinations. The judges ranked the scripts in each pack in the order of their relative quality.

Analysis of grade boundary data

All judges completed the tasks successfully. The analytical methods for determining grade boundaries were different for traditional awarding on the one hand, and for Thurstone pairs and rank ordering methods on the other. All judgements from the traditional awarding task were sent to the appropriate Chairs of Examiners, who were asked to look at the judges' decisions and determine the zones of uncertainty and grade boundaries for each judge group.

For the rank ordering data, FACETS software (Linacre, 2005) was used to employ multi-faceted Rasch analysis, which allowed scripts from 2006 and 2007 to be placed on the same scale of perceived quality. The raw mark scales of the two examinations could then be compared directly so that mark x in one year could be deemed equivalent to mark y in the other year in terms of perceived quality of candidate performance.

For Thurstone pairs, Rasch analysis was also employed. However, due to the very restricted mark ranges of the scripts used, (which were very close to the grade boundaries), it was inappropriate to directly relate the mark scale to the scale of perceived quality in this case. We therefore used a crude method of calculating the equivalent marks, which used the following formula:

2007 Thurstone implied boundary =

2007 mean mark - [(SD 2007 mark/SD 2007 measure)

× (Mean 2007 measure – Mean 2006 measure)].

The boundary marks generated by the Thurstone pairs task therefore have to be viewed with some caution.

Findings relating to grade boundaries

The grade boundary marks for 2007 that were generated experimentally by the three methods are summarised in Tables 2 and 3 (biology), and 4 and 5 (English).

For biology, intra-method reliability was excellent for traditional awarding: the boundary marks generated were identical across the three judge groups for one boundary, and identical for two judge groups on the other boundary. The reliability of Thurstone pairs was also very high: for both grade boundaries, the boundary marks were identical for two judge groups, while the boundary mark of the third group differed by only one mark. The intra-method reliability of rank ordering was slightly lower but still very high: it was perfect for the A/B grade boundary, but for the E/U boundary three different boundary marks were generated, all one mark apart.

For English, the findings were similar. Although for four of the six boundaries to be determined, the Chair of Examiners felt unable to complete the task without referring to statistical indicators, the zones of uncertainty restricted potential grade boundaries to such an extent that it was still possible to conclude that the intra-method reliability of traditional awarding was high. The intra-method reliability of the

2 In a linked study, the judges also completed a fourth task in which they rated scripts on a number of different features. This was part of a wider research project, presented by Novaković and Suto (2009).

Table 2: Summary of E/U grade boundary marks for biology

Task	*Judge group*			*Actual 2007 grade boundary mark*
	Group 1	*Group 2*	*Group 3*	
Traditional awarding	16	16	16	17
Thurstone pairs	15	16	15	
Rank ordering	14	14	14	

Table 3: Summary of A/B grade boundary marks for biology

Task	*Judge group*			*Actual 2007 grade boundary mark*
	Group 1	*Group 2*	*Group 3*	
Traditional awarding	35	34	34	34
Thurstone pairs	33	33	32	
Rank ordering	32	31	33	

Table 4: Summary of C/D grade boundary marks for English

Task	*Judge group*			*Actual 2007 grade boundary mark*
	Group 1	*Group 2*	*Group 3*	
Traditional awarding	56	55	? (54–56)	55
Thurstone pairs	55	55	56	
Rank ordering	56	57	58	

Table 5: Summary of A/B grade boundary marks for English

Task	*Judge group*			*Actual 2007 grade boundary mark*
	Group 1	*Group 2*	*Group 3*	
Traditional awarding	? (69–70)	? (69–70)	? (68–70)	69
Thurstone pairs	69	70	69	
Rank ordering	69	68	72	

Thurstone pairs method was also very high. For both grade boundaries, two groups generated the same boundary mark, whereas the mark of the third group differed by only one mark. Intra-method reliability was again lower for rank ordering. For the C/D grade boundary, three different boundary marks were generated, all one mark apart. For the A/B grade boundary, all three boundary marks were different, and spanned a five-mark range.

There was no overall trend in leniency/severity across the judge groups for either subject: no single group generated boundary marks that were consistently higher or lower than the marks of the other two groups. This finding may be taken to confirm that the judge groups in the study were well matched.

When the three methods are compared with one another, it appears that for both subjects, the traditional awarding and Thurstone pairs methods generated very similar boundary marks, except for the biology A/B grade boundary. The boundary marks generated by rank ordering were all on the lenient side for biology, whereas for the English C/D grade boundary, they were on the severe side. However, without using additional research methods to triangulate findings, it is not possible to determine which of these June 2007 grade boundary marks are equivalent ontologically to the actual 2006 boundary marks. It is therefore not possible to conclude from this study which method, if any, is ultimately the most effective at maintaining standards.

Limitations

While the research has found traditional awarding to have high intra-method reliability, there is a possibility that this reliability is simply an artefact of the method – even if the 'zone' had been as wide as the mark range, it is possible that the boundary mark would still have been chosen in the middle. A possible way of investigating intra-method reliability of traditional awarding in more detail would be to give different groups of examiners scripts covering non-identical mark ranges (offset by a few marks) and ask them to set the grade boundaries. In our study, however, we wanted to keep the procedure as close as possible to the one used at live awarding meetings.

One of the major limitations relates to the way that the Thurstone pairs method was used in our study, that is, for the purpose of producing grade boundaries. As there is no existing procedure for using Thurstone pairs as a grading method, we used it as it has been used in comparability studies, using scripts only in a small range around the grade boundary. This made it impossible to calculate equivalent marks by plotting pairs of regression lines (as in rank-ordering), and the grade boundary marks for this method were calculated using an alternative and rather crude method. These marks therefore need to be regarded with caution.
A better way of using Thurstone pairs for grading purposes would be to use the scripts covering a wide mark range, although this might prove impractical or tiring, considering the number of judgements that would need to be made. Kimbell *et al*. (2007) have been using Thurstone pairs for grading purposes on a wide range of marks; however, they have used Thurstone pairs in combination with rank-ordering (thus creating a hybrid grading method), and they have so far not proposed a way of translating the experts' judgements into the actual grades.

A limitation of all three methods is their reliance on particular individuals for critical judgements. For traditional awarding, the zones of uncertainty and grade boundaries were judged by Chairs of Examiners alone, as it was impractical for them to harness the other judges' collective professional judgement. For Thurstone pairs and rank ordering, the researchers made equally crucial judgements during the Rasch analyses, about which misfitting or outlying scripts and judgements to exclude.

Conclusions

It can be concluded from this study that, reassuringly, none of the three methods investigated is strikingly weak in terms of either type of reliability, and all three methods appear to have functioned well, generating highly plausible grade boundaries. Whilst theoretically, methods that rely on comparative rather than absolute judgements might be favourable (Laming, 2004), this study provides no empirical evidence to support such a preference. The implication of this is that any of the three methods explored could contribute to the determination of grade boundaries operationally.

Table 6: Final comparison of traditional awarding, Thurstone pairs and rank ordering

	Traditional awarding		*Thurstone pairs*		*Rank ordering*	
	Biology	*English*	*Biology*	*English*	*Biology*	*English*
Intra-method reliability	Excellent	Very high	Very high	Very high	Very high	Reasonable
Inter-method reliability	Quite high with Thurstone pairs and rank ordering	Very high with Thurstone pairs; quite high with rank ordering	Quite high with traditional awarding and rank ordering	Very high with traditional awarding; quite high with rank ordering	Quite high with traditional awarding and Thurstone pairs	
Number of judgements made per judge in ½ a day's work	20	24	40	40	40	40
Key operational advantages	No need for scripts to be cleaned of marks.		Requires no extra input from the Chair of Examiners.		Requires no extra input from the Chair of Examiners.	
Key operational disadvantages	Requires considerable input from Chair of Examiners.		Scripts must be cleaned of marks (less problematic for scripts marked on-screen). Requires a large quantity of archive scripts.		Scripts must be cleaned of marks (less problematic for scripts marked on-screen). Requires a large quantity of archive scripts.	
Key theoretical strengths	Draws on the collective expertise of 'communities of practice', though only while meetings continue to be largely face to face. Arguably, remote awarding risks weakening these communities.		Relies on relative rather than absolute judgements. Unaffected by judges' leniency or severity.		Relies on relative rather than absolute judgements. Unaffected by judges' leniency or severity. Large number of paired comparisons obtained from actual human judgements.	
Key theoretical weaknesses	Relies on absolute rather than relative judgements. Affected by judges' leniency or severity. Judgements are 'contaminated' by statistical information.		Rasch techniques (e.g. FACETS) are often used to analyse the data – the modelling assumption of a single latent trait is controversial. The mark range covered by scripts is too small to calculate equivalent marks without making considerable assumptions. When scripts cover a wide mark range, the judges' task can become tiresome, and rank-ordering lends itself better to producing a large number of comparisons.		Rasch techniques (e.g. FACETS) are often used to analyse the data – the modelling assumption of a single latent trait is controversial. Places significant demands on the working memory.	

Overall, the results of our study do not provide enough evidence to favour one method over the other two, either for operational or research purposes. However, in Table 6 we have drawn together the findings from our study and from other research and anecdotal evidence relating to the three methods. We hope it will prove useful to anyone making a decision about which method to use. It is important to emphasise once again that while rank ordering has been used for grading purposes previously, there is no existing procedure for using Thurstone pairs in determining grade boundaries. In this table, we have listed the advantages and disadvantages of Thurstone pairs as it has been used in this study (adapted from comparability studies).

References

Bell, J.F., Bramley, T. & Raikes, N. (1998). Investigating A level mathematics standards over time. *British Journal of Curriculum and Assessment* **8**, 2, 7–11.

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, 27–29 August in Berlin, Germany.

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

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

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

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

Elliott, G., Johnson, N. & Bramley, T. (2005). *Cross-validation of 2004 standard setting in GCE AL Psychology 2540 using a rank-ordering methodology*. Cambridge Assessment internal report.

Gill, T. & Black, B. (2006). *An investigation of standard maintaining and equating using expert judgment in GCSE English between years and across tiers using a rank-ordering method*. Cambridge Assessment internal report.

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 Conference, 5–8 September in London, UK.

Gill, T. & Bramley, T. (2008). *How accurate are examiners' judgments of script quality?* Paper presented at the British Educational Research Association Annual Conference, 3–6 September in Edinburgh UK.

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

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*. Paper presented at the British Educational Research Association Conference, 5–8 September in London, UK.

Greatorex, J. & Nádas, R. (2008). *Using 'thinking aloud' to investigate judgements about A-level standards: does verbalising thoughts result in different decisions?* Paper presented at the British Educational Research Association Annual Conference, 3–6 September in Edinburgh, UK.

Kimbell, R., Wheeler, A., Miller, S. & Pollitt, A. (2007). *E-scape portfolio assessment phase 2 report*. Technology Education Research Unit, Goldsmiths College, University of London.

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

Linacre, J.M. (2005). FACETS Rasch measurement computer program. www.winsteps.com

Novaković, N. & Suto, I. (2009). *How should grade boundaries be determined in examinations? An exploration of the script features that influence expert judgements*. Paper presented at the European Conference on Educational Research, 28–30 September in Vienna, Austria.

OCR. (2008). *OCR Procedures for Awards*. Revised April 2008. Cambridge: OCR.

Raikes, N., Scorey, S. & Shiell, H. (2008). *Grading examinations using expert judgements from a diverse pool of judges*. Paper presented at the 34th Annual Conference of the International Association for Educational Assessment, 7–12 September in Cambridge, UK.

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

Thurstone, L.L. (1927b). A law of comparative judgment. *Psychological Review*, **3**, 273–286.

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

How do examiners make judgements about standards? Some insights from a qualitative analysis

Jackie Greatorex Research Division

An earlier version of this article was presented at the American Educational Research Association conference, April 2009

Abstract

There is a good deal of research about how judgements are made in awarding when A level and GCSE grade boundaries are chosen. There is less research about how judgements are made in Thurstone paired comparisons and rank ordering (popular methods in comparability studies to compare grading standards). Therefore, the research question for the present study is 'how do Principal Examiners[1] (PEs) make judgements about standards in awarding, Thurstone paired comparisons and rank ordering?' The present article draws from a wider project in which Principal Examiners thought aloud whilst making judgements about the quality of candidates' work and grading standards in awarding, Thurstone paired comparisons and rank ordering situations analogous to how these methods are practised. For the present analysis a coding frame was developed to qualitatively analyse the think aloud data. The coding frame constituted codes grounded in the think aloud data and grade descriptors[2] from the qualification specification. It was found that overall the Principal Examiners attended to valid factors such as where marks were gained, responses to key questions and characteristics of candidates' work that were in the grade descriptors. When the importance of each factor was considered there were some similarities and some differences between the methods. Implications and recommendations are discussed.

Background

The focus of this article is the often asked question 'how do Principal Examiners make judgements about standards?' This question can be addressed from various perspectives including:

- What cognitive strategies do PEs use?
- What features do PEs attend to (and are they valid features)?
- What procedures are used to make decisions?

In the current article three approaches to judging grading standards are considered: (i) awarding – part of the conventional approach to recommending grade boundaries, (ii) Thurstone pairs and (iii) rank ordering. The latter two were suggested as possible future methods of

1 Principal Examiners generally write an examination question paper, lead the associated marking and take part in awarding. Most participants in Thurstone paired comparison and rank ordering studies are Principal Examiners.

2 Grade descriptors (descriptions) are written descriptions that indicate the level of attainment characteristic of a particular qualification. They give a general indication of the learning outcomes at a given grade. The descriptions should be interpreted in relation to the content outlined in specifications, they do not outline the specification content (OCR, 2004). A specification is a description of what can be tested in an examination. Note that this research was undertaken before specifications began providing **performance descriptions** rather than **grade descriptions**. Performance descriptors (descriptions) are written descriptions of the typical knowledge, skills and understanding likely to be found in candidates' work at the judgementally awarded grade boundaries. These descriptors are indicators of the knowledge, understanding and skills that are likely to be found in candidates' work at the grade boundary, they are not requirements. There might be other knowledge, understanding and skills that are found in candidates' work at the grade boundary. They are designed to aid recommending grade boundaries.