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.

recommending grade boundaries by Pollitt and Elliott (2003a and b), Black and Bramley (2008) and Kimbell *et al*. (2007). They have also been used in a series of comparability studies (e.g. Forster and Gray, 2000; Arlett, 2003; Greatorex *et al*., 2002, 2003; Edwards and Adams, 2002, 2003; Guthrie, 2003; Bramley *et al*., 1998; Townley, 2007). Note that Thurstone pairs and rank ordering are not currently used in operational awarding or in operational procedures to recommend grade boundaries.

What are the current practices for awarding, Thurstone paired comparisons and rank ordering?

In this research the focus is on one decision-making phase of awarding which involves the awarding committee judging whether a small number of examples of candidates' work[3] on particular marks show the distinguishing characteristics of performance at a particular grade. For a fuller description, see Cresswell (1997), QCA (2008) or Greatorex (2003a).

Thurstone pairs and rank ordering have been frequently described in the literature and there are many examples of their use in comparability studies; see for example Bramley *et al*. (1998), Arlett (2003), Greatorex *et al*. (2002, 2003), Edwards and Adams (2002, 2003), Guthrie (2003) and Townley (2007). Both Thurstone pairs and rank ordering involve a group of experts judging the quality of candidates' work.

In a Thurstone pairs design each expert compares a pair of scripts. In a study investigating standards maintenance, each pair would consist of a script from the most recent examination and one from the archive examination. Each expert decides which of two scripts contains the better performance, without re-marking the scripts. This is repeated for a variety of pairs of scripts. Once all the necessary comparisons are complete, they are statistically analysed (using Rasch). The results of the analysis can be used to identify a small range of marks within which the most recent boundary should lie for the standard from last year to be maintained.

In a study investigating standards maintenance using a rank ordering design each expert is given small samples of scripts which they rank from best to worst performance. Each small sample has a mixture of most recent and archive scripts. This is repeated for a number of overlapping samples of scripts. The outcomes of the rankings are submitted to the same statistical analysis as above. Again the statistics can be used to identify a small range of marks within which the most recent boundary should lie.

What does research tell us about how judgements are made about grading standards?

There is a good deal of research about judgements of grading standards in awarding, for example, Good and Cresswell (1988a and b), Scharaschkin and Baird (2000), Baird and Scharaschkin (2002). The present literature review will be confined to the most relevant literature.

Murphy *et al*. (1995) argue that each awarding committee member's impressions of what was appropriate were from a variety of sources, three of which were identified in their research:

1. knowledge of requirements of the national curriculum or other descriptions of performance;
2. performance on questions that some believed to be indicative of achievement (and the belief that it was possible to make judgements on these alone);
3. the belief that they 'knew' what constitutes work at a particular grade.

They found that the general use of archive material was low. Later Baird (2000) found that the severity of judgements of grade-worthiness was sometimes influenced by the archives provided. Research shows archive scripts were sometimes missed in awarding in the past. Archive scripts are still a useful source of information listed in the Code of Practice.

Cresswell (1997) investigated the weighting of many factors in judgements about grading standards such as technical and statistical evidence as well as the features noted in candidates' work. He found little evidence that the demand of questions was taken into account when PEs judged the candidates' work. Cresswell (1997) and later Crisp (2007, 2008) found that valid features of candidates' work contributed to decisions about grading standards. Crisp (2008) found that PEs made judgements by paying attention to features in candidates' work which were closely tied to the mark scheme, such as a good understanding of concepts, application of knowledge and evaluation and application of skills. However, Cresswell (1997) also argued that other less valid features also had some input in judgements of grading standards. For example, sometimes features such as whether the candidate's work gave the reader pleasure or was interesting were taken into account, when they were not necessarily linked to the features intended to be judged (Cresswell, 1997).

There are various aspects of awarding meetings and scripts that positively and negatively influence judgements of gradeworthiness (Cresswell, 1997; Murphy *et al*., 1995; Crisp, 2007; Baird, 2000; Baird and Scharaschkin, 2002; Scharaschkin and Baird, 2000). To consider this further it is important to note that A level and GCSE examinations have a principle of compensation, according to which candidates gain marks for their strengths, and there is more than one way to achieve a grade. Two conundrums relate to the principle of compensation and the visibility of marks on scripts:

- Some PEs in some awarding meetings particularly focus on questions and marks which are believed to differentiate between performances at particular grades (Murphy *et al*., 1995; Greatorex *et al*., 2008). This belief might be well or ill founded (Murphy *et al*., 1995). Focussing on particular questions at the expense of other questions is not aligned with the principle of compensation. Psychological research from a variety of contexts suggests that humans are not particularly good at combining information to make decisions. For a detailed discussion of this, see Greatorex (2007) and Greatorex *et al*. (2008). Therefore, focussing judgements on particular questions might be a successful approach to decision making, if the questions are a good proxy for the whole of the examination. After all, the other strategy – judgements about whole scripts – involves mentally combining a candidate's answers to all questions in the examination.
- It has been established that the consistency of candidates' performance across questions on an examination paper influences the severity of judgements of gradeworthiness (Cresswell, 1997; Scharaschkin and Baird, 2000). Again, this is not aligned with the principle of compensation.

3 The candidates' work is usually written examination scripts but might be a recording of a drama or musical performance or an artefact such as a painting.

There is a small amount of research about how judgements are made in Thurstone paired comparisons comparability studies. For example, Edwards and Adams (2002, 2003) asked PEs in Thurstone paired comparison studies what criteria they used to judge the candidates' work. They report that the criteria were quite wide ranging, but that some of the common criteria included "depth of understanding" and "level of reasoning" (Edwards and Adams 2003, p.20). All the examples that they list seem to be valid and reasonable criteria for judging the candidates' work. This reassures us that for some Thurstone paired comparison studies judgements are made by taking valid information into account. In rank ordering studies the correlation between the trait 'perceived quality of candidates' work' and total mark is pleasingly high (between 0.8 and 0.9) (Bramley, 2007). Thus we have some evidence that rank ordering is measuring something similar to the total marks, and that the judgements are valid.

Context of the present study

The present study is the third in a series of inter-linked studies which draw from a wider research project. The research is still in progress. The first and second studies are reported in Greatorex and Nádas (2009) and Greatorex *et al*. (2008). In the wider project the aim is to find out more about cognitive processes used by PEs to make judgements about grading standards.

Greatorex and Nádas (2009) found that, broadly speaking, the task outcomes were similar whether the judgements were made silently or whilst thinking aloud. Therefore, there is some evidence that research results using the think aloud data are trustworthy.

Greatorex *et al*. (2008) studied which examination question responses or answers the PEs referred to in the candidates' work. They found that the questions most often referred to did not always discriminate well between achievements just above and below the grade boundary. This ties in with Murphy *et al*.'s (1995) concern that the questions used as key discriminators might or might not statistically discriminate between performances on the two adjacent grades. Therefore, the Research Division at Cambridge Assessment argued that question level data from on-screen marking should be used to facilitate choosing key discriminating questions.

Thus far the reporting of the wider research project, of which this study forms a part, has covered a quantitative analysis of the outcomes of the tasks, and qualitative coding using *a priori* codes (the examination questions). What has not been reported is a qualitative analysis using codes that are grounded in the rich textual content of the think aloud data, and therefore this is the focus of the present study.

Method

The method for the project is reported in more detail in Greatorex and Nádas (2009). Two past AS biology examinations were used as a source of data. The first year of the examination will be referred to as the 'archive examination' and the next year of the examination will be referred to as the 'live examination'. The five participants (called PEs in this report) had all been involved in awarding the AS examination. All the examples of candidates' work used in the research were from near the grade boundaries from the two examinations.

Prior to the main data collection phase PEs undertook some warm up exercises including:

- Thinking aloud whilst doing non-examining tasks.
- Silently making decisions in the five experimental conditions described below.

In the main data collection phase PEs thought aloud whilst making judgements in the five experimental conditions:

- Awarding with marks visible ('awarding visible');
- Awarding with candidates' work cleaned of marks ('awarding clean');
- Thurstone paired comparisons with marks visible ('Thurstone visible');
- Thurstone paired comparisons with candidates' work cleaned of marks ('Thurstone clean');
- 'Rank ordering' with candidates' work cleaned of marks.

The thinking aloud was audio recorded and transcriptions were made.

The awarding conditions reflected the aspect of awarding where individual committee members evaluate scripts, before coming to a collective view about where the grade boundary should be. The rank ordering and Thurstone pairs conditions were intended to reflect current/best practices in prior studies. For all experimental conditions some small adjustments were made to current/best practices for the purposes of this research. Photocopies of the scripts were used rather than the original scripts. For each method the scripts were presented as they are normally presented: awarding with marks visible and rank ordering with scripts cleaned of marks. Thurstone pairs studies vary regarding whether the marks are visible or not so this was reflected in the research. 'Awarding clean' reflected the aspect of awarding where individual awarding committee members evaluate scripts, before coming to a collective view about where the grade boundary should be. But in 'awarding clean' the scripts were cleaned of marks. A reason for this experimental control was the arguably extraneous influence of visible marks in some awarding judgements (Murphy *et al*., 1995; Cresswell, 1997; Scharaschkin and Baird, 2000).

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

Coding for the present study

The present study involved developing a coding frame to qualitatively analyse the think aloud data. The coding frame constituted codes grounded in the think aloud data and grade descriptors from the qualification specification. To develop the coding frame the transcripts, instructions to PEs, examples of candidates' work and grade descriptors were read. Although the grade descriptors are not used in the grading process, it is likely that they would give a good indication of senior examiners' views of achievement at each grade. Over a series of iterations of reading and trying out codes and coding frames, a coding frame grounded in the data was developed. The process was informed by some of the content of the transcripts as well as anecdotal conversations with the PEs.

The final coding frame is described in Table 1 and Table 2. Some codes were used to identify when examiners paid attention to responses to

Table 1: The coding frame of codes grounded in the think aloud data and the question papers

Shorthand label used in coding the transcripts	*What the question(s) required candidates to do/topics tested*	*What the PEs seem to be doing*
'Archive/ Question A'	Explain the significance of the different affinities of foetal haemoglobin and adult haemoglobin for oxygen	The PEs seemed to consider this question to be a high demand question and therefore a good source of information about A and B grade performance.
'Archive/ Question B'	Explain the significance of the dissociation curve of adult haemoglobin	As above
'Comparing long answers'	Explain the relationship between the structure and function of arteries, veins and capillaries	One question in each examination was a long answer question on a somewhat similar topic so sometimes the answers from different years were compared or referred to.
'Live/Question X'	Explain translocation as an energy requiring process	The PEs seemed to consider this question to be a high demand question and therefore a good source of information about A and B grade performance.
'Live/Question X–'	Explain translocation as an energy requiring process	The PEs seemed to consider this question to be a high demand question and therefore a good source of information about A and B grade performance. This code applied only to negative comments about the candidates' work.
'Live/Question X+'	Explain translocation as an energy requiring process	The PEs seemed to consider this question to be a high demand question and therefore a good source of information about A and B grade performance. This code applied only to positive comments about the candidates' work.
'Live/Question Y'	Describing the mammalian circulatory system as a closed double circulation	Question Y in the live examination was arguably a lower demand question than those listed above but seems to have been seen as a good source of information.

Table 2: The coding framework of codes grounded in the think aloud data and the mark scheme or grade descriptors

Shorthand label used in coding the transcripts	*What the PE seems to be doing*
'Explain'	The PE seems to be looking for a characteristic listed in the grade descriptor, i.e. provide coherent and logical explanations.
'Identify marks'	The PE seems to be trying to identify where marks were given.
'Know and understand'	The PE seems to be looking for a characteristic listed in the grade descriptor, i.e. show good knowledge and understanding.
'Present'	The PE seems to be looking for a characteristic listed in the grade descriptor, i.e. present ideas clearly and logically.

particular items, these are given in Table 1. Other codes were grounded in the protocols, mark scheme and grade descriptors (see Table 2). Each code was taken to be a factor that contributed to judgements about grading standards.

Unfortunately, for some PEs there was not time to complete all the tasks and in places transcripts are ambiguous, resulting in some missing data.

A sample of data was double-coded. The second coder did not see the original coding. Once the double-coding was collated, only the most reliably coded codes were retained.

Once the coding was complete, it was established which code(s) was present in the section of the transcript associated with each example of candidate's work. Next, the presence data was expressed as a proportion of the total number of examples of candidates' work available in each condition for all PEs. For instance, the following is a hypothetical example: there were 100 examples of candidates' work in total in 'rank ordering'. Code A was present for candidates 1 to 60, so code A had a proportion of 60%, whereas, code B was present only for candidates 5 and 6, and so had a proportion of 2%. The proportions were ranked in descending order. Therefore, the higher the rank, the more important the code (or associated factor) is in making judgements. Using our example the factor associated with code A was more important in making judgements than the factor associated with code B. A limitation of this analysis is that some information is lost by ranking rather than using frequencies or similar.

Results

Overall, the PEs made judgements in all the conditions by paying attention to:

- Responses about particular areas of content (questions) which seemed to be perceived as a good source of information about A and B grade performance and/or were perceived to be high demand questions.
- Responses to the long answer question in each examination which had some overlap in the subject content tested, and therefore seemed a solid basis for comparison between the performance in the two different examinations.
- Some characteristics referenced in the grade descriptors.
- Whether the candidates seemed to have been credited with marks.

This is summarised in Figure 1.

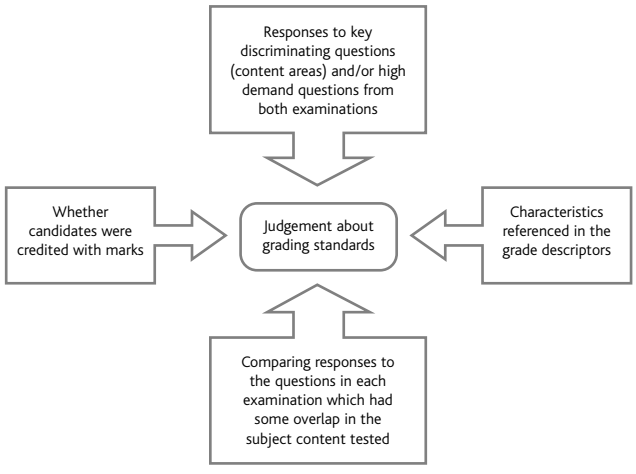

Figure 1: The overarching themes that contributed to judgements

In addition to the overarching themes that contributed to judgements about grading standards there were the factors identified in the coding frame. The following text boxes give the rank of the importance of each factor in judgements for each condition. Note that some of the ranks are ties and therefore some ranks are repeated and others are omitted. For example, for 'awarding visible' two codes were ranked 9 and no codes were ranked 10.

How were judgements made in 'awarding visible'?

Overall judgements were made by looking for correct answers, focussing on answers to particular questions or areas of the syllabus and looking for characteristics listed in grade descriptors. In descending order of importance the factors taken into account were:

1 Responses to question Y in the live examination. This question was about the mammalian circulatory system and seemed to be considered a good source of information about A/B grade performance.

2 Identifying where marks were given.

3 Comparing responses to the long answer questions (one in each examination) which were both about explaining the relationship between the structure and function of arteries, veins and capillaries. These questions seemed to be perceived as a good source of information about grade A/B performance.

4 Negative views about performance on question X in the live examination. This question expected candidates to explain translocation as an energy requiring process.

5 Positive views about performance on question X in the live examination.

6 Responses to question B in the archive examination. This question expected candidates to explain the significance of the dissociation curve of adult haemoglobin and seemed to be viewed as a good source of information about grade A/B performance.

7 Finding evidence of a characteristic in the grade descriptors; in this case 'presents ideas clearly and logically'.

8 Neutral views about performance on question X in the live examination.

9 Finding evidence of characteristics in the grade descriptors; in this case 'provides coherent and logical explanations'.

9 Finding evidence of characteristics in the grade descriptors; in this case 'shows good knowledge and understanding'.

11 Responses to question A in the archive examination. This question expected candidates to explain the significance of the different affinities of foetal haemoglobin and adult haemoglobin for oxygen and seemed to be viewed as a good source of information about grade A/B performance.

How were judgements made in 'awarding clean'?

Overall judgements were made by looking for correct answers, focussing on answers to particular questions and looking for characteristics listed in grade descriptors. In descending order of importance the factors taken into account were:

1 Responses to question Y in the live examination. This question was about the mammalian circulatory system and seemed to be considered a good source of information about A/B grade performance.

2 Identifying where marks were given.

3 Comparing responses to the long answer questions (one in each examination) which were both about explaining the relationship between the structure and function of arteries, veins and capillaries. These questions seemed to be perceived as a good source of information about grade A/B performance.

4 Positive views about performance on question X in the live examination. This question expected candidates to explain translocation as an energy requiring process.

5 Responses to question A in the archive examination. This question expected candidates to explain the significance of the different affinities of foetal haemoglobin and adult haemoglobin for oxygen and seemed to be viewed as a good source of information about grade A/B performance.

6 Negative views about performance on question X in the live examination.

7 Finding evidence of a characteristic in the grade descriptors; in this case 'provides coherent and logical explanations'.

8 Responses to question B in the archive examination. This question expected candidates to explain the significance of the dissociation curve of adult haemoglobin and seemed to be viewed as a good source of information about grade A/B performance.

9 Neutral views about performance on question X in the live examination.

10 Finding evidence of a characteristic in the grade descriptors; in this case 'presents ideas clearly and logically'.

11 Finding evidence of characteristics in the grade descriptors; in this case 'shows good knowledge and understanding'.

How were judgements made in 'Thurstone clean'?

Overall judgements were made by looking for correct answers, focussing on answers to particular questions and looking for characteristics listed in grade descriptors. In descending order of importance the factors taken into account were:

1 Identifying where marks were given.

2 Comparing responses to the long answer questions (one in each examination) which were both about explaining the relationship between the structure and function of arteries, veins and capillaries. These questions seemed to be perceived as a good source of information about grade A/B performance.

2 Responses to question Y in the live examination. This question was about the mammalian circulatory system and seemed to be considered a good source of information about A/B grade performance.

4 Responses to question A in the archive examination. This question expected candidates to explain the significance of the different affinities of foetal haemoglobin and adult haemoglobin for oxygen and seemed to be viewed as a good source of information about grade A/B performance.

5 Responses to question B in the archive examination. This question expected candidates to explain the significance of the dissociation curve of adult haemoglobin and seemed to be viewed as a good source of information about grade A/B performance.

6 Finding evidence of a characteristic in the grade descriptors; in this case 'provides coherent and logical explanations'.

7 Positive views about performance on question X in the live examination. This question expected candidates to explain translocation as an energy requiring process.

7 Negative views about performance on question X in the live examination.

9 Neutral views about performance on question X in the live examination.

10 Finding evidence of a characteristic in the grade descriptors; in this case 'shows good knowledge and understanding'.

11 Finding evidence of a characteristic in the grade descriptors; in this case 'presents ideas clearly and logically'.

How were judgements made in 'Thurstone visible'?

Overall judgements were made by looking for correct answers, focussing on answers to particular questions and looking for characteristics listed in grade descriptors. In descending order of importance the factors taken into account were:

1 Identifying where marks were given.

2 Comparing responses to the long answer questions (one in each examination) which were both about explaining the relationship between the structure and function of arteries, veins and capillaries. These questions seemed to be perceived as a good source of information about grade A/B performance.

2 Responses to question A in the archive examination. This question expected candidates to explain the significance of the different affinities of foetal haemoglobin and adult haemoglobin for oxygen and seemed to be viewed as a good source of information about grade A/B performance.

4 Responses to question B in the archive examination. This question expected candidates to explain the significance of the dissociation curve of adult haemoglobin and seemed to be viewed as a good source of information about grade A/B performance.

5 Responses to question Y in the live examination. This question was about the mammalian circulatory system and seemed to be considered a good source of information about A/B grade performance.

6 Finding evidence of a characteristic in the grade descriptors; in this case 'presents ideas clearly and logically'.

7 Negative views about performance on question X in the live examination. This question expected candidates to explain translocation as an energy requiring process.

7 Positive views about performance on question X in the live examination.

9 Finding evidence of a characteristic in the grade descriptors; in this case 'shows good knowledge and understanding'.

10 Neutral views about performance on question X in the live examination.

11 Finding evidence of a characteristic in the grade descriptors; in this case 'provides coherent and logical explanations'.

How were 'rank ordering' judgements made?

Overall judgements were made by looking for correct answers, focussing on answers to particular questions and looking for characteristics listed in grade descriptors. In descending order of importance the factors taken into account were:

1 Identifying where marks were given.

2 Responses to question A in the archive examination. This question expected candidates to explain the significance of the different affinities of foetal haemoglobin and adult haemoglobin for oxygen and seemed to be viewed as a good source of information about grade A/B performance.

3 Comparing responses to the long answer questions (one in each examination) which were both about explaining the relationship between the structure and function of arteries, veins and capillaries. These questions seemed to be perceived as a good source of information about grade A/B performance.

4 Responses to question B in the archive examination. This question expected candidates to explain the significance of the dissociation curve of adult haemoglobin and seemed to be viewed as a good source of information about grade A/B performance.

5 Responses to question Y in the live examination. This question was about the mammalian circulatory system and seemed to be considered a good source of information about A/B grade performance.

6 Negative views about performance on question X in the live examination. This question expected candidates to explain translocation as an energy requiring process.

7 Finding evidence of a characteristic in the grade descriptors; in this case 'provides coherent and logical explanations'.

7 Positive views about performance on question X in the live examination.

9 Neutral views about performance on question X in the live examination.

10 Finding evidence of a characteristic in the grade descriptors; in this case 'presents ideas clearly and logically'.

11 Finding evidence of characteristics in the grade descriptors; in this case 'shows good knowledge and understanding'.

There are some commonalities in the importance of the different factors in the judgements made in different conditions (see Table 3). 'Identify marks' was ranked amongst the two most important factors for all conditions, and 'comparing long answers' was in the top three most important factors for all conditions. Additionally, 'know and understand' (showing good knowledge and understanding) was ranked amongst the three least important factors for all conditions. 'Live/question X' was also ranked amongst the four least important factors for all conditions.
There were also some differences in the rank order of importance of the different factors in different conditions (see Table 3). Factor 'archive/question A' was ranked in the top five most important factors for the 'awarding clean', 'rank ordering', 'Thurstone visible' and 'Thurstone clean' conditions, but was ranked as the least important factor for 'awarding visible'. Factor 'live/question Y' was ranked in the top two most important factors for the 'awarding clean', 'awarding visible' and 'Thurstone clean' conditions, but was ranked as lower for the 'rank ordering' and 'Thurstone visible' conditions. Factor 'live/question X+' was ranked as fourth most important for the 'awarding clean' condition but as low as seventh for the 'rank ordering', 'Thurstone clean' and the 'Thurstone visible' conditions. The factor 'present' was ranked as low (tenth or lower) for the 'Thurstone clean', 'awarding clean' and 'rank ordering' conditions, but seventh or higher for the 'awarding visible' and 'Thurstone visible' conditions.

Discussion

The main research question for the present study is 'how do Principal Examiners make judgements about grading standards in awarding, Thurstone paired comparisons and rank ordering?' It was found that overall the PEs attended to valid factors such as where marks were gained, responses to key questions and characteristics of candidates' work that are referenced in the grade descriptors. This finding was apparent for all conditions, and might be somewhat generalisable to the methods – awarding, Thurstone paired comparisons and rank ordering. When the importance of each factor was considered there were some similarities and some differences between the methods.

There are a number of limitations to the present study. First, it is not possible to generalise about all GCSE and A-level judgements of grading standards from two examinations and one judgementally awarded grade

Table 3: The rank order of importance of each factor in judgements of grading standards

Shorthand label	*rank order of importance in judgements*				
	'awarding visible'	*'awarding ' clean*	*'rank ordering'*	*'Thurstone visible'*	*'Thurstone clean'*
'Archive/question A'	11	5	2	2 =	4
'Archive/question B'	6	8	4	4	5
'Comparing long answers'	3	3	3	2 =	2 =
'Live/question X'	8	9	9	10	9
'Live/question X–'	4	6	6	7 =	7 =
'Live/question X+'	5	4	7 =	7 =	7 =
'Live/question Y'	1	1	5	5	2 =
'Explain'	9 =	7	7 =	11	6
'Identify marks'	2	2	1	1	1
'Know and understand'	9 =	11	11	9	10
'Present'	7	10	10	6	11

Note: 1 is the highest rank; = denotes ties

boundary. However, the examinations were carefully chosen as examinations which might involve judgements about numerical skills, written skills, use of diagrams, and knowledge and understanding, whereas in some other subjects PEs might judge candidates' work which is predominantly in one skill area. Secondly, only think aloud was used as a method of data collection. It is often advised that think aloud data are used to generate hypotheses which are tested out in further empirical studies. To this end there is research underway at Cambridge Assessment to identify which features of candidates' work are used in judgements about grading standards using a more quantitative and generalisable approach. Thirdly, the 'awarding clean' and 'awarding visible' conditions have limited ecological validity; they do not include much of the information that is available in traditional awarding meetings, and they omit the face to face social dynamics of the awarding meeting. For research that incorporates these influences see Murphy *et al*. (1995) and Cresswell (1997). However, the awarding meeting information was not provided to avoid it influencing the judgements in the other conditions. Furthermore, if remote awarding becomes more widespread then there might be an increase in individual decision making which reflects the think aloud setting in this study when a PE made judgements without other PEs present.

The general themes that the PEs attended to (characteristics referenced in the grade descriptors, key discriminating questions, comparing answers to similar questions from different years of the examination and identifying where marks were given) all seem to be valid sources of information for making judgements about grading standards. The limitations and strengths of using key discriminating questions have been considered by Murphy *et al*. (1995) and later by Greatorex *et al*. (2008). For example, more credit might be given to responses to particular questions than was intended by the mark scheme. Additionally, it is important that the question is measuring the same as the whole examination. Comparing answers to similar questions from the two years of the examination shares the strengths and limitations of using key discriminating questions. The finding that PEs attend to some specific items, and that the items seem to be used because of the demands they place on candidates, illustrates that the context in which the candidates perform is important to PEs' decision making. This is a contrast to Cresswell's finding that the PEs did not pay much attention to the demands of the questions and how this affected candidates' performance.

Much of the previous literature has suggested that PEs compare the candidates' work with their impression of what is appropriate to a particular grade (sometimes referred to as a prototype or internal standard) (Murphy, 1995; Baird, 2000; Crisp, 2008). In the present analysis it was found that PEs attended to features referenced in the grade descriptors. In line with current awarding practices the grade descriptors were not available during the thinking aloud and therefore the PEs must have been remembering them, or the descriptors are a good reflection of the prototypes that PEs have for performance at grades A and B. This ties in with the well-rehearsed argument that grade descriptors should be grounded in both candidates' actual performance and Principal Examiners' views of the features that discriminate between achievement at different grades (Greatorex, 2001, 2002, 2003b; Greatorex *et al*., 2001). PEs seem to be looking for particular features and using particular features in judgements whether they are comparing the candidates' performance with a prototype, or with a memory of another candidate's work.

Crisp (2007) and Bramley (2007) indicate that there is commonality between what is given credit in the mark scheme (measured by total mark) and what contributes to judgements of grading standards. This ties in with the finding in the present analysis that PEs try to identify what marks were given.

The general themes which contributed to judgements of grading standards reflect some of the existing literature. However, what has not previously been reported is a comparison of the judgement process in awarding versus Thurstone paired comparisons versus rank ordering, and this is the focus of the next section.

There were some commonalities between the factors that were ranked as the most and least important factors in making judgements. For instance, 'comparing long answers' was ranked high for all conditions, and this corroborates the findings of Greatorex *et al*. (2008). Therefore, it seems that there are some commonalities in how PEs make judgements in each of the conditions. On the other hand there were also some differences in the rank of importance of the different factors in different conditions. There was no clear overall pattern regarding whether two or more conditions were particularly similar in how PEs made judgements.

There were some differences in the rank order of importance of the various factors in different conditions. The factor 'present' was ranked as low (tenth or lower) for the 'Thurstone clean', 'awarding clean' and 'rank ordering' conditions, but seventh or higher for the 'awarding visible' and 'Thurstone visible' conditions. Also 'archive/question A' and 'archive question B' were ranked lower in 'awarding visible' and in 'awarding clean', than in the comparability study conditions. This appears to somewhat corroborate Murphy *et al*'s (1995) finding that the archive scripts are infrequently used, however, awarding practices have changed since their work and the Code of Practice (2008, p36) says that the archive "must be used, as appropriate, to inform the determination of marks at key grade boundaries". Indeed Laming's (2004) work about humans being better at making comparisons than maintaining internal standards would suggest that as far as possible awarding procedures should recommend systematic

and frequent comparisons between the archive and the live examples of candidates' work. It is not clear why the importance of some other factors varies between conditions. For example, 'live/question Y' is amongst the two most important factors in the 'awarding visible', 'awarding clean' and 'Thurstone clean' but is of lower ranking in the other conditions.

Previous research has tended to compare the trait measured in comparability studies with total marks rather than awarding judgements; see for example Bramley (2007). However, the present study offers the opportunity to compare what might be measured in comparability studies with what is measured in awarding. This is accomplished by treating what PEs attend to as a strong proxy for what is measured. The present study suggests the trait 'perceived quality of candidates' work' might vary a little with the condition that is used in comparability studies (rank ordering or Thurstone paired comparisons), and might also differ somewhat from what is measured in awarding at a particular boundary. However, as explained earlier there are also strong commonalities between conditions regarding both the factors PEs attend to and their importance in judgements. If there were system changes as suggested by Pollitt and Elliott (2003a and b) or Black and Bramley (2008) then what is being measured might change slightly. However, in all approaches in this research PEs attended to valid factors, so what was measured when using each method is arguably valid.

The present study has offered many insights into what PEs attend to when they make the judgements about grading standards, from psychological and other perspectives. However, it is somewhat difficult to generalise from this particular analysis to other examinations, as some of the coding refers to aspects of biology. The next stage in the wider research project is to undertake a more psychological analysis with particular focus on whether PEs are making comparisons between candidates' work or whether they are using internal standards.

Acknowledgements

Thank you to the senior examiners who took part in this research. Thank you also to Richard Shewry who provided confidential consultancy for this research project; amongst other things he offered insights from the biology teaching and assessment community which is not the specialist area of the main researcher from this project.

References

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

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

Baird, J. & Scharaschkin, A. (2002). Is the whole worth more than the sum of the parts? Studies of examiners' grading of individual papers and candidates' whole A-level examination performances. *Educational Studies*, **28**, 2, 143–162.

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

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

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

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

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

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

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

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

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

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

Good, F. J. & Cresswell, M. J. (1988a). Grade Awarding Judgements in differentiated examinations. *British Educational Research Journal*, **14**, 3, 263–281.

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

Greatorex, J. (2001). Making the grade – how question choice and type affect the development of grade descriptors. *Educational Studies*, **27**, 4, 451–464.

Greatorex, J. (2002). Making Accounting examiners' tacit knowledge more explicit: developing grade descriptors for Accounting A-Level. *Research Papers in Education*, **17**, 2, 211–226.

Greatorex, J. (2003a). *What happened to limen referencing? An exploration of how the Awarding of public examinations has been and might be conceptualised*. A paper presented at the British Educational Research Association Conference, Heriot Watt University, Edinburgh, September 2008.

Greatorex, J. (2003b). Developing and applying level descriptors. *Westminster Studies in Education*, **26**, 2, 125–133.

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

Greatorex, J. & Nádas, R. (2009). Using 'thinking aloud' to investigate judgements about A-level standards: does verbalising thoughts result in different decisions? *Research Matters: A Cambridge Assessment Publication*, **7**, 8–16. Also presented at the British Educational Research Association Conference, Heriot Watt University, Edinburgh, September 2008.

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

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

Greatorex, J., Johnson, C. & Frame, K. (2001). Making the grade – developing grade profiles for accounting using a discriminator model of performance. *Westminster Studies in Education*, **24**, 2, 167–181.

Greatorex, J., Novaković, N. & Suto, I. (2008). *What attracts judges' attention? A comparison of three grading methods*. A paper presented at the International Association for Educational Assessment Conference, Cambridge, September 2008.

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

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

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

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

OCR (2004). OCR AS GCE Business Studies (3811) OCR Advanced GCE in Business Studies (7811) Approved Specification, Revised Edition. OCR. [online.] Available at: http://www.ocr.org.uk/qualifications/as_alevelgce/business_studies/documents.html

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

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

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

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

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

'Key discriminators' and the use of item level data in awarding

Tom Bramley Research Division

Introduction

As more examination papers in general qualifications (GCSEs and A levels) are scanned and marked on screen, the marks on individual questions or question parts are collected automatically, and are referred to as item level data (ILD). The analysis of ILD is available for use in awarding meetings (where the grade boundaries are decided). This article discusses the theoretical rationale for using ILD in awarding, presents some possible formats for displaying data, and suggests ways in which the data could be used in practice.

For many examinations (whether marked on screen or not), the Principal Examiner (PE) will have produced a list of the questions which they expected to be 'key discriminators' at particular grade boundaries. This information might come from the test blueprint (for example, if each question on a test was 'targeted' at pupils at a particular grade or level), or it might come from the PE's (and their marking team's) experience of marking the papers – for example, if during the course of marking the paper they noticed which questions seemed to be discriminating well at particular grades or levels.

The (often unspoken) assumption behind identifying these 'key discriminators' is that by focussing on performance on these questions when making judgements about scripts in the awarding meeting, the awarding panel will use their time and effort most efficiently and be best able to identify the overall score on the test which represents the same performance standard as the corresponding grade boundary set in previous sessions.

The Guttman pattern – an idealised scenario

Imagine that we have a test consisting of ten dichotomous items (items scored 1 or 0). The scores on such a test fit a Guttman[1] pattern if success on an item implies success on all easier items and failure on an item implies failure on all harder items. If the columns represent the items with the easiest item at the left and the hardest item at the right, and the rows represent examinees with the least able at the top and the most able at the bottom, then a Guttman pattern for scores of 23 examinees on this 10-item test might look like Table 1 below.

If the score data fit this idealised pattern then all scripts on the same test total would show exactly the same performance (in terms of which items were answered correctly and incorrectly). In other words, every script perfectly represents the performance of all examinees with the same test score. Furthermore, there is a 'simple order' in the raw scores. Each increasing test total implies that the examinee has achieved

1 Louis Guttman (1916–1987) was an American psychologist. See http://en.wikipedia.org/wiki/Guttman_scale for more information.