# How are archive scripts used in judgements about maintaining grading standards?

Jackie Greatorex

Cambridge Assessment

**A paper presented at the British Educational Research Association Annual Conference, September 2009, Manchester**

**Contact:**
Jackie Greatorex
Core Research Group
Research Division
Cambridge Assessment
1 Hills Road
Cambridge
CB1 2EU
Direct dial. 01223 553835
Fax. 01223 552700
Email: greatorex.j@cambridgeassessment.org.uk

# How are archive scripts used in judgements about maintaining grading standards?

# ABSTRACT

Generally GCE and GCSE Awarding Bodies use:

- *Awarding* procedures to determine grade boundaries.
- *Comparability studies* to monitor standards over time or between Awarding Bodies.

Both Thurstone pairs and rank ordering involve judging the quality of scripts, and are used in some comparability studies. Pollitt and Elliott (2003a and b), Pollitt (2004) and Kimbell *et al* (2007) suggested replacing **much of marking and awarding** with *Thurstone pairs,* whereas Black and Bramley (2008) suggested replacing **the script evaluation aspect of awarding** with *rank ordering*, in the GCSE, AS and A-level context. Neither Thurstone pairs nor rank ordering are currently used to determine candidates' GCSE, AS and A-level results. These ideas are still being explored, refined and debated.

The article has two aims:

- To compare Thurstone pairs, rank ordering and the script evaluation aspect of awarding in terms of the archive items receiving most attention.
- To identify how well these archive items discriminated between the performance of candidates who received grades A and B.

Concurrent think aloud verbal protocols were collected as five senior examiners judged script quality in different experimental conditions. The conditions included replicating:

- Thurstone pairs
- Rank ordering
- The script evaluation aspect of awarding

The procedures were replicated as closely as possible within the confines of the study.

Results indicated that:

1. Examiners referenced some items relatively more frequently than others and the most referenced items varied with condition
2. Two items statistically discriminated between item level marks of candidates who were awarded grades A and B
3. These two items were not always the most referenced items

The first result suggests that slightly different constructs are measured in each condition, which is somewhat at odds with some previous research findings. Research to confirm the constructs used to make comparisons in Awarding, Thurstone pairs and rank ordering is still ongoing (King *et al*, 2009). These investigations do not rely on verbal protocol data.

The second and third results:

- are in line with previous research indicating that subject experts have varied success in predicting item level statistics including how well different groups of students will perform (Cadwell, 1950; Impara and Plake, 1998; Hambleton and Jirka, 2006);
- support the use of item level data in Awarding to help Awarders focus their judgements on appropriate items.

# INTRODUCTION

## *Background*

Employers and education institutions use GCSE and GCE A-level qualification results, along with other information, to select people for employment or additional study. These GCSE and A-level results might be from various Awarding Bodies and/or different years. Therefore, it is important that there is comparability of standards between Awarding Bodies and over time.

There are a number of approaches to determining *grade boundaries*[i]*;* they involve either *standard setting* or *maintaining a standard from one examination session to the next*. Descriptions of some standard setting methods can be found in texts such as Angoff (1987) or Hambleton and Pitoniak (2006). There are various approaches to maintaining standards from one examination session to another, and for investigating the comparability of examinations or qualifications. Many of the contemporary UK methods for investigating the comparability of standards are given in Newton *et al.* (2007). Some methods of standard setting and/or maintaining standards are predominantly statistically orientated whilst others have a larger role for subject experts' judgements of the quality of examinees' performances (scripts[ii]). Judging the quality of *archive scripts*[iii] and maintaining standards are the key issues addressed in this article.

Three approaches to maintaining standards over time are considered: (i) the script evaluation aspect of awarding, (ii) Thurstone pairs and (iii) rank ordering. Awarding is the conventional approach to recommending GCSE and A-level grade boundaries. Thurstone pairs and rank ordering have 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; Black and Bramley, 2008). Pollitt and Elliott (2003a and b), Pollitt (2004) and Kimbell *et al* (2007) suggested replacing **much of marking and awarding** with *Thurstone pairs,* whereas Black and Bramley (2008) suggested replacing **the script evaluation aspect of awarding** with *rank ordering.* Neither Thurstone pairs nor rank ordering are currently used to determine candidates' GCSE, AS and A-level results. All three approaches involve subject experts judging scripts. The focus of this paper is how the subject experts use archive scripts in their judgements.

## *What are current awarding, Thurstone pairs and rank ordering practices?*

Generally, the awarding procedure is designed to maintain standards from one examination session to another. Only in the case of a new qualification is awarding a matter of standard setting. Mostly the awarding process is undertaken by an *Awarding Committee* (Senior Examiners or awarders). This article considers one decision-making phase of awarding which involves judging script quality. QCA explains that:

"Examiners do this by looking closely at a sample of scripts with marks near where each boundary is likely to be. They have to judge which are worthy of an A (or an E) and which are not. It is clearly important that the judgements they make are consistent with previous examinations and between Awarding Bodies. To do this, examiners use a variety of materials, including archive scripts from previous examinations" (www.qca.org.uk/qca_7006.aspx).

This process is repeated for each judgementally determined grade and *unit*[iv]. Once the judgmentally determined grade boundaries are decided they are used arithmetically to calculate the remaining grade boundaries. In an *operational*[v] context there are also other checks and balances in addition to the Awarding Committee, including the Accountable Officer authorising the grade boundaries.

For over two decades archive scripts on grade boundaries have been used in Awarding Committee meetings to focus examiners on the previous year's grade boundary standard and enable it to be maintained for operational grade boundaries. At least one Awarding Body has considered replacing these archives with an *anchor archive* from a particular year of a new *specification* (syllabus) and

this archive would not be replenished for the duration of the specification. An anchor archive should represent a stable standard when stakeholders are familiar with the specification.

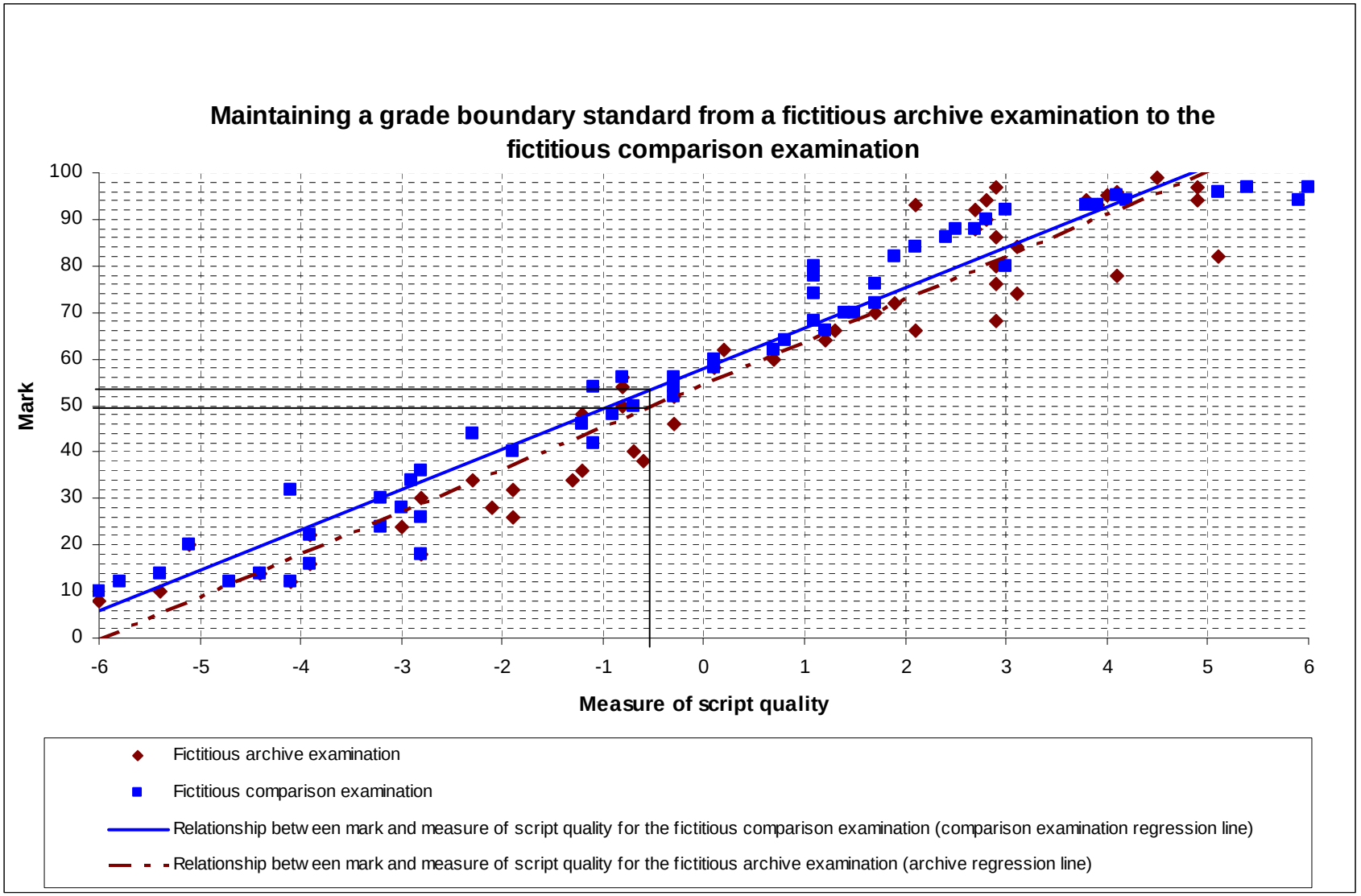
For a fuller description about awarding see Cresswell (1997), QCA (2008) or Greatorex (2003).

Thurstone pairs and rank ordering as well as examples of their use in comparability studies have been frequently described in the literature; see for example Bramley *et al.* (1998), Arlett (2003), Greatorex *et al.* (2002, 2003), Edwards and Adams (2002, 2003), Guthrie (2003), Townley (2007) and Bramley (2007). Therefore, a summary tailored towards using Thurstone pairs or rank ordering to choose a grade boundary is provided[vi]. Thurstone pairs and rank ordering involve a group of experts judging the quality of candidates' work. The group of experts are generally the examiners from the relevant Awarding Committees. In Thurstone pairs each expert compares a pair of scripts, with each pair constituting a script from the archive examination and second script from a comparison examination. Each expert decides which of the scripts shows evidence of better candidate performance, without re-marking the scripts. This is repeated for a variety of pairs of scripts. When all the necessary comparisons have been made, they are statistically analysed by fitting a Rasch model. The Rasch analysis places all the scripts on a scale measuring perceived script quality. The results of the analysis can be used to identify where the comparison examination boundary should lie for the standard from last year to be maintained.

Rather than using data from an empirical study the following illustration of how a Rasch analysis might be used to maintain standards utilises fictitious examinations and fictitious data. In the case of Figure 1 the measure of script quality was plotted against the marks on two fictitious examinations (archive examination and comparison examination). The highest total mark available for each fictitious examination was 100 marks. Linear regression lines showing the relationship between mark and measure of script quality for each fictitious examination were plotted. To find where the comparison examination boundary should be based on the archive grade boundary the graph should be used as follows. If the archive grade boundary was 50 marks, it corresponds with the measure of script quality -0.5 because these values intersect on the archive regression line. This same measure of script quality intersects with the comparison examination regression line at about 54 marks. In this case the grade boundary for the comparison examination would be 54 marks.

In rank ordering each expert receives small 'packs' of archive and comparison examination scripts which they rank order according to the quality of the candidates' performance. This is repeated for a number of packs. The judgements are analysed by fitting a Rasch model. The results of the Rasch analysis can be used in the same way as described for Thurstone pairs.

**Figure 1**


Maintaining a grade boundary standard from a fictitious archive examination to the fictitious comparison examination
Mark
100
90
80
70
60
50
40
30
20
10
0
-6
-5
-4
-3
-2
-1
0
1
2
3
4
5
6
Measure of script quality
Fictitious archive examination
Fictitious comparison examination
Relationship betw een mark and measure of script quality for the fictitious comparison examination (comparison examination regression line)
Relationship betw een mark and measure of script quality for the fictitious archive examination (archive regression line)

There are a number of aspects of awarding meetings and scripts that positively and negatively influence judgements of gradeworthiness (Murphy *et al.*, 1995; Cresswell, 1997; Baird, 2000; Scharaschkin and Baird, 2000; Baird and Scharaschkin, 2002; Crisp, 2007). Arguably, in judgements of gradeworthiness the visibility of marks given to candidates' responses can become extraneous information. For instance, some Awarding Committee members pay particular attention to items 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). Focusing judgements on particular items might be a successful approach to decision making, if the items are a good proxy for the whole of the examination. If the item is not a good proxy for the whole examination then arguably the marks for this item can be an extraneous variable in decision making. Additionally, it has been established that the consistency of candidates' performance across items on an examination paper influences the severity of judgements of gradeworthiness (Cresswell, 1997; Scharaschkin and Baird, 2000). Again this is an example of how marks can be an extraneous variable in decisions. In some Thurstone pairs studies marks are not visible and for other Thurstone pairs studies the marks are visible. In rank ordering studies the marks are not visible. As yet the influence of the visibility of marks on Thurstone pairs and rank ordering decisions has not been established.

## *What does research tell us about the use of archive scripts in awarding, Thurstone pairs and rank ordering?*

Murphy *et al*. (1995) researched several aspects of awarding using observations, questionnaires and interviews. They stated that

"the general use of archive materials was low" (Murphy *et al.*, 1995, 33).

Cresswell (1997) also researched awarding practices and the cognitive process used in judging the quality of operational scripts in awarding. There was no mention of the term *archive script* in his thesis, and his research put forward scarce evidence for the use of archive scripts. These studies are quite old, and in the decade since the studies were completed practices have evolved. The research literature offers scant evidence about what examiners attend to in archive scripts. For example, Murphy *et al.* (1995) did not indicate what features of the archive scripts were salient in decision making in awarding. Additionally, Cresswell (1997) and Crisp (2007) focused on the features examiners attended to in scripts during awarding, but Cresswell did not distinguish between live[vii] and archive scripts, whereas Crisp's work did not include archive scripts.

Baird (2000) conducted significant research about the use of archive scripts in judgements of grading standards. In her experiment there were two subjects (Psychology and English Literature), each with four groups of examiners matched for severity or leniency of judgements of gradeworthiness. These four groups of examiners received different archive information:

- One group received no archive scripts
- A second group received archive scripts on the operational grade E boundary with balanced mark profiles
- A third group received archive scripts on the operational grade E boundary with unbalanced mark profiles
- A fourth group received archive scripts which were on the operational grade D boundary (the paper implied that the examiners thought the scripts were on the grade E boundary)

The results for the different experimental groups were compared. The manipulation of archive scripts affected the judgements in Psychology but not English Literature. Therefore, some examiners used their personal view of grade boundary standards and other examiners relied on archive scripts to focus them on the grade boundary standard.

There is little public domain research about judging archive scripts in Thurstone pairs or rank ordering. This is partly because many of the studies utilising Thurstone pairs were about standards between Awarding Bodies in the same year; so no archive was necessary.

### *Aims*

This article draws from a wider project which is work in progress. The aim of the wider project is to investigate the psychology of the decision-making processes used in Thurstone pairs, rank ordering and judging scripts in awarding. This article has two subsidiary aims:

- To compare Thurstone pairs, rank ordering and the script evaluation aspect of awarding in terms of the archive items receiving most attention.
- To identify how well these archive items discriminated between the performance of candidates who received grades A and B.

Only details from the wider project which are relevant to this article are presented. For more details about the wider project see Greatorex *et al* (2008), Greatorex (2009) and Greatorex and Nadas (2009).

## METHOD

### *Design*

There were two stages to the research. Initially the participants made awarding, Thurstone pairs and rank ordering judgements silently (i.e. without being asked to verbalise their thinking). This was partly to familiarise them the comparability study techniques. The tasks were then repeated whilst thinking aloud as the main data collection phase.

### *Examination*

Two past AS-level science examinations from the same qualification and adjacent years were used. One examination is referred to as the *live* examination and the other as the *archive* examination.

In the operational examination the question papers were given to candidates in a form in which the items and source material (e.g. diagrams) were presented along with an answer space into which they added their responses.

### *Scripts*

The scripts had total marks within the range of marks considered in the recommendation for the grade A boundary in the operational awarding meeting. Photocopies of the scripts were used rather than the original scripts. Different scripts were used in the silent tasks and in the verbal protocol tasks.

### *Participants*

Five senior examiners who were members of the operational Awarding Committee for the AS-level examinations took part in the research.

### *Conditions*

Overall there were five experimental conditions in which participants made judgements about grading standards whilst:

- Awarding with marks visible ('*awarding visible*')
- Awarding with candidates' work cleaned of marks ('*awarding clean*')
- Thurstone pairs with marks visible ('*Thurstone pairs visible*')
- Thurstone pairs with candidates' work cleaned of marks ('*Thurstone pairs clean*')
- '*Rank ordering*' with candidates' work cleaned of marks

For all conditions, the question paper, scripts and mark scheme were available for reference.

*Awarding visible* and *awarding clean* reflected the aspect of awarding where individual Awarding Committee members evaluated scripts, before coming to a collective view about where the grade boundary should be. The other conditions reflected Thurstone pairs and rank ordering study practices with any required adjustments for the purposes of this study.

For the main data collection phase the aim was for each participant to experience the conditions one after the other in the Cambridge offices with a researcher present. The participants were provided with instructions for each condition which reflected usual practices; they were asked to think out loud as they undertook each condition. Additionally, they were asked to say which script and item they were "looking at or thinking about". If the participant was quiet for some time then the researcher asked them to "please keep talking".

The verbal protocols were digitally recorded with the permission of the participants. Subsequently, the digitally recorded information was transcribed.

## *Guarding against order effects*

Various precautions were taken to guard against order effects in the main data collection phase. These included breaks or distractor tasks between conditions, the careful allocation of scripts to conditions and participants as well as the varying order in which the participants experienced the conditions.

## *Analysis*

In this article results of the analysis of the archive examination data are given.

The items in the archive examination were renumbered using Roman numerals so that 1ai became I, 1aii became II and so on for all items.

Initially the transcripts were coded for references to items and scripts for the thinking aloud conditions. For each condition and associated script, the appearance or nonappearance of one or more references to each item was noted. The proportion of available archive scripts on which each item was referenced was calculated. Next the items were ranked according to their relative frequencies of referencing (for each condition), e.g., the rank used for the most frequently referenced items was 1.

In previous studies researchers have used and interpreted verbal protocol data in various ways including:

- Frequently referenced (mentioned) information was considered to be more important in judgements than infrequently referenced information
- When research participants judged performance the frequently referenced information was considered to be a stronger proxy for the construct being measured than the infrequently referenced information

For a more detailed discussion of verbal protocol analysis as well as examples of studies using verbal protocols see Ericsson and Simon (1980, 1993), Green (1998), Cushing-Weigle (1999), Backlund *et al* (2003), Crisp (2007), Suto and Greatorex (2008).

Following the practices outlined above:

- Relatively frequently referenced items were interpreted to contribute more to judgements than the relatively infrequently referenced items
- Relatively frequently referenced items were also interpreted to be a stronger proxy for what was measured in each condition than relatively infrequently referenced items.

The items were classified into different types based on the kind of answer the candidates were expected to provide, using the system devised by Rita Nádas and used previously in Greatorex *et al.* (2008). The classifications used were: explain, 1-2 words, calculation, gap filling, one number, labelling and long answer.

Subsequently an item analysis using marks given during operational marking was undertaken for

the archive scripts. These scripts included both the scripts that were used in the silent tasks and the scripts which were used in the main data collection phase. The analysis identified items that distinguished between the achievement of grade A and grade B candidates from the operational examination results. Mann Whitney U tests were used to compare the rank of item marks from grade B candidates with the rank of item marks of grade A candidates. If there was a statistically significant difference between grade A and grade B candidates' marks on an item then this item was thought to provide sound evidence for making decisions about grading standards, as long as the item was a good proxy for what is measured by the examination.

# RESULTS

Table 1 indicates: a) ranking of items in order of their relative frequency of referencing, b) the number of archive scripts for which participants referenced each item and c) the number of archive scripts for which participants referenced an item as a percentage of the total number of available archive scripts in the thinking aloud conditions.

## *Referenced items*

Information about the importance of each item in decision making is given in table 1. Item XIII the long answer item, was ranked 1 or 2 for all conditions. Item IV was ranked 2 for four conditions, items I, VI and VII were ranked 1 or 2 for three conditions, items VIII and XI were ranked 1 or 2 for two conditions, and items III, XII and XV were ranked 1 or 2 for one condition. Therefore, there were differences in the importance of each item in decision making between conditions, and each condition measured something a little different to the others.

## *Referenced item types*

The relatively frequently referenced items were classified as long answer (item XIII), explain (items I, IV, VII, VIII, XI, XII, XV), calculation and gap filling (item III), one number and calculation (item VI). The item types which were relatively less frequently referenced were: 1-2 words (item II), explain (items V, XIV, XVI, XVII, XVIII) and labelling (items IX, X).

## *For which items are there statistically significant differences between the item level marks of grade A and B candidates?*

There was a statistically significant difference between the mean marks of grade A and grade B candidates for items VIII and XII (see table 2). Therefore these items were considered to statistically discriminate between marks of the two groups at the item level. The mean mark of grade A candidates was higher than the mean mark of grade B candidates on these items (see table 2). The item analysis justified the participants' relatively frequent referencing of item VIII in *rank ordering* and *Thurstone pairs visible*, as well as the relatively frequent referencing of XII in *awarding clean*. Generally the participants did not reference the statistically discriminating items in decision making. Both items VIII and XII required an explanation from the candidate.

**Table 1**: a) Ranking of items in order of their relative frequency of referencing, b) Number of archive scripts for which participants referenced an item, and c) Number of archive scripts for which participants referenced an item as a percentage of available archive scripts in the thinking aloud conditions.

| | | | a) Rank in order of relative frequency of referencing (importance in decision making) | | | | | b) No. of archive scripts for which participants referenced an item | | | | | c) No. of archive scripts for which participants referenced an item as a % of the available scripts in the thinking aloud conditions. | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Item** | Maximum marks available | What is needed from the candidate? | *awarding visible* | *awarding clean* | *rank ordering* | *Thurstone pairs visible* | Thurstone pairs | *awarding visible* | *awarding clean* | Rank ordering | *Thurstone pairs visible* | *Thurstone pairs clean* | *awarding visible* | *awarding clean* | Rank ordering | *Thurstone pairs visible* | *Thurstone pairs clean* |
| I | 4 | Explain | 2 | 2 | 3 | 5 | 1 | 3 | 4 | 14 | 15 | 20 | 37.50 | 50.00 | 70.00 | 75.00 | 100.00 |
| II | 1 | 1-2 words | 4 | 3 | 9 | 12 | 7 | 1 | 3 | 6 | 8 | 12 | 12.50 | 37.50 | 30.00 | 40.00 | 60.00 |
| III | 1 | Calculation, Gap filling | 1 | 4 | 11 | 11 | 8 | 4 | 2 | 0 | 9 | 11 | 50.00 | 25.00 | 0.00 | 45.00 | 55.00 |
| IV | 3 | Explain | 2 | 2 | 2 | 4 | 2 | 3 | 4 | 17 | 17 | 19 | 37.50 | 50.00 | 85.00 | 85.00 | 95.00 |
| V | 2 | Explain | 4 | 5 | 5 | 6 | 3 | 1 | 1 | 11 | 14 | 17 | 12.50 | 12.50 | 55.00 | 70.00 | 85.00 |
| VI | 1 | 1 number, Calculation | 2 | 2 | 2 | 8 | 5 | 3 | 4 | 17 | 12 | 15 | 37.50 | 50.00 | 85.00 | 60.00 | 75.00 |
| VII | 4 | Explain | 3 | 2 | 3 | 2 | 1 | 2 | 4 | 14 | 19 | 20 | 25.00 | 50.00 | 70.00 | 95.00 | 100.00 |
| VIII | 2 | Explain | 3 | 3 | 1 | 1 | 3 | 2 | 3 | 19 | 20 | 17 | 25.00 | 37.50 | 95.00 | 100.00 | 85.00 |
| IX | 2 | Labelling | 3 | 4 | 9 | 13 | 6 | 2 | 2 | 6 | 7 | 14 | 25.00 | 25.00 | 30.00 | 35.00 | 70.00 |
| X | 1 | Labelling | 3 | 4 | 10 | 14 | 6 | 2 | 2 | 4 | 6 | 14 | 25.00 | 25.00 | 20.00 | 30.00 | 70.00 |
| XI | 2 | Explain | 3 | 1 | 4 | 3 | 2 | 2 | 5 | 12 | 18 | 19 | 25.00 | 62.50 | 60.00 | 90.00 | 95.00 |
| XII | 2 | Explain | 4 | 2 | 6 | 3 | 4 | 1 | 4 | 10 | 18 | 16 | 12.50 | 50.00 | 50.00 | 90.00 | 80.00 |
| XIII | 6 | Long answer | 1 | 2 | 2 | 2 | 1 | 4 | 4 | 17 | 19 | 20 | 50.00 | 50.00 | 85.00 | 95.00 | 100.00 |
| XIV | 2 | Explain | 4 | 5 | 6 | 9 | 3 | 1 | 1 | 10 | 11 | 17 | 12.50 | 12.50 | 50.00 | 55.00 | 85.00 |
| XV | 4 | Explain | 4 | 5 | 7 | 10 | 2 | 1 | 1 | 9 | 10 | 19 | 12.50 | 12.50 | 45.00 | 50.00 | 95.00 |
| XVI | 2 | Explain | 3 | 4 | 8 | 10 | 7 | 2 | 2 | 8 | 10 | 12 | 25.00 | 25.00 | 40.00 | 50.00 | 60.00 |
| XVII | 4 | Explain | 3 | 5 | 7 | 7 | 4 | 2 | 1 | 9 | 13 | 16 | 25.00 | 12.50 | 45.00 | 65.00 | 80.00 |
| XVIII | 1 | Explain | 4 | 5 | 10 | 12 | 7 | 1 | 1 | 4 | 8 | 12 | 12.50 | 12.50 | 20.00 | 40.00 | 60.00 |

Note that in total there were 8 archive scripts available in *awarding visible*, i.e. 4 participants each had 2 archive scripts available. The same was true for *awarding clean*. In total there were 20 archive scripts available in *Thurstone pairs visible*, i.e. 4 participants each had 5 archive scripts. The same was true for *Thurstone pairs clean* and *rank ordering*. Due to time constraints verbal protocols from all participants for all five conditions were not available for analysis. It was not always the same four participants whose protocols were analysed for each condition.

**Table 2:** a) Mann Whitney U tests to compare the rank or item marks from grade B candidates with the rank of item marks of grade A candidates, b) Descriptive statistics for grade A candidates, and c) Descriptive statistics for grade B candidates.

| | a) Candidates who gained a grade A or B | | | | | | | | | b) Candidates who gained a grade A | | | | | c) Candidates who gained a grade B | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Item | U | Significance (exact) | Significance level | maximum marks available | mean | sd | n | min | max | mean | sd | n | min | max | mean | sd | n | min | max |
| I | 63.50 | 0.151 | > 0.05 | 4 | 3.31 | 0.81 | 29 | 2 | 4 | 3.47 | 0.77 | 19 | 2 | 4 | 3.00 | 0.82 | 10 | 2 | 4 |
| II | 86.00 | 0.701 | > 0.05 | 1 | 0.86 | 0.35 | 29 | 0 | 1 | 0.89 | 0.32 | 19 | 0 | 1 | 0.80 | 0.42 | 10 | 0 | 1 |
| III | 85.50 | 0.668 | > 0.05 | 1 | 0.97 | 0.19 | 29 | 0 | 1 | 1.00 | 0.00 | 19 | 1 | 1 | 0.90 | 0.32 | 10 | 0 | 1 |
| IV | 64.50 | 0.164 | > 0.05 | 3 | 1.69 | 0.93 | 29 | 0 | 3 | 1.53 | 0.90 | 19 | 0 | 3 | 2.00 | 0.94 | 10 | 0 | 3 |
| V | 95.00 | 1.000 | > 0.05 | 2 | 1.41 | 0.63 | 29 | 0 | 2 | 1.42 | 0.61 | 19 | 0 | 2 | 1.40 | 0.70 | 10 | 0 | 2 |
| VI | 92.50 | 0.910 | > 0.05 | 1 | 0.52 | 0.51 | 29 | 0 | 1 | 0.53 | 0.51 | 19 | 0 | 1 | 0.50 | 0.53 | 10 | 0 | 1 |
| VII | 79.00 | 0.484 | > 0.05 | 4 | 2.72 | 0.84 | 29 | 1 | 4 | 2.79 | 0.92 | 19 | 1 | 4 | 2.60 | 0.70 | 10 | 2 | 4 |
| **VIII** | **51.50** | **0.045** | **< 0.05** | 2 | 0.76 | 0.69 | 29 | 0 | 2 | 0.95 | 0.62 | 19 | 0 | 2 | 0.40 | 0.70 | 10 | 0 | 2 |
| IX | 83.00 | 0.604 | > 0.05 | 2 | 1.76 | 0.51 | 29 | 0 | 2 | 1.79 | 0.54 | 19 | 0 | 2 | 1.70 | 0.48 | 10 | 1 | 2 |
| X | 81.00 | 0.542 | > 0.05 | 1 | 0.90 | 0.31 | 29 | 0 | 1 | 0.95 | 0.23 | 19 | 0 | 1 | 0.80 | 0.42 | 10 | 0 | 1 |
| XI | 71.50 | 0.286 | > 0.05 | 2 | 1.86 | 0.35 | 29 | 1 | 2 | 1.95 | 0.23 | 19 | 1 | 2 | 1.70 | 0.48 | 10 | 1 | 2 |
| **XII** | **49.00** | **0.035** | **< 0.05** | 2 | 1.52 | 0.51 | 29 | 1 | 2 | 1.68 | 0.48 | 19 | 1 | 2 | 1.20 | 0.42 | 10 | 1 | 2 |
| XIII | 90.00 | 0.839 | > 0.05 | 6 | 5.17 | 1.04 | 29 | 3 | 6 | 5.21 | 1.03 | 19 | 3 | 6 | 5.10 | 1.10 | 10 | 3 | 6 |
| XIV | 88.00 | 0.769 | > 0.05 | 2 | 1.72 | 0.53 | 29 | 0 | 2 | 1.68 | 0.58 | 19 | 0 | 2 | 1.80 | 0.42 | 10 | 1 | 2 |
| XV | 69.00 | 0.247 | > 0.05 | 4 | 3.45 | 0.69 | 29 | 2 | 4 | 3.32 | 0.75 | 19 | 2 | 4 | 3.70 | 0.48 | 10 | 3 | 4 |
| XVI | 87.00 | 0.735 | > 0.05 | 2 | 1.66 | 0.48 | 29 | 1 | 2 | 1.68 | 0.48 | 19 | 1 | 2 | 1.60 | 0.52 | 10 | 1 | 2 |
| XVII | 64.50 | 0.164 | > 0.05 | 4 | 3.21 | 0.86 | 29 | 1 | 4 | 3.37 | 0.83 | 19 | 1 | 4 | 2.90 | 0.88 | 10 | 2 | 4 |
| XVIII | 78.00 | 0.456 | > 0.05 | 1 | 0.52 | 0.51 | 29 | 0 | 1 | 0.58 | 0.51 | 19 | 0 | 1 | 0.40 | 0.52 | 10 | 0 | 1 |

Note these statistics include scripts from the warm up exercises as well as the main data collection phase.

# LIMITATIONS

The first limitation was that although the *rank ordering*, *Thurstone pairs clean* and *Thurstone pairs visible* conditions were very similar to current/best practices, *awarding visible* and *awarding clean* slightly digressed from operational awarding practice in two ways:

(i) Participants in the experiment did not have some of the information that would usually be available in operational awarding (apart from the archive scripts, question paper and the mark scheme). This was to avoid influencing the decisions made in the other conditions, which do not include using such information.

(ii) Participants in the experiment made individual decisions. Operationally awarders do not always make decisions individually about the quality of candidates' work, although individual decision making is not uncommon (Cresswell, 1997). Individual rather than collaborative decisions about individual scripts might increase *if* the practice of undertaking awarding meetings remotely[viii] becomes more widespread.

Therefore, *awarding visible* and *awarding clean* might have somewhat limited ecological validity.

A second limitation was that the robustness of the statistics was compromised by the small samples of participants and scripts which affected the generalisability of the study. Despite these limitations there were some important results.

# DISCUSSION

Item XIII was relatively frequently referenced (ranked 1 or 2 for all conditions). Item IV was relatively frequently referenced (ranked 2 for four conditions). Item XIII required a long answer and candidates could score a maximum of 6 marks on this item. Item IV required an explanation from the candidate and candidates could score a maximum of 3 marks on this item. Relatively frequently referenced items were interpreted to contribute more to judgements than the relatively infrequently referenced items. Therefore, overall items XIII and IV contributed more to judgements than the other items.

There were other items in addition to XIII and IV which were frequently referenced. Items I, VI and VII were ranked 1 or 2 for three conditions, items VIII and XI were ranked 1 or 2 for two conditions, and items III, XII and XV were ranked 1 or 2 for one condition. Therefore, there were differences in the importance of each item in decision making between conditions, and each condition measured something slightly different to the others. Therefore if the approach to aspects of awarding or marking were replaced with Thurstone pairs and or rank ordering then what is measured might change slightly. Additionally, the choice of *Thurstone pairs visible* or *Thurstone pairs clean* or *rank ordering* to conduct a comparability study might make a slight difference to the measure used to compare grading standards. This contrasts with the findings of previous research, which show that the measure of script quality from Thurstone pairs studies and rank ordering studies correlate well with total mark, for example see Bramley (2007). Such correlations suggest that studies using either method measure a similar trait to total mark. The debate is still inconclusive and research is still underway to contribute to the accumulation of evidence. For instance, King *et al* (2009) report a study in which Awarding, Thurstone pair and rank ordering judgements, along with ratings of scripts for different features were collected. One line of investigation will aim to confirm the constructs used to make judgements and comparisons in each method. This will be accomplished by statistically linking the ratings of script features with Awarding, Thurstone pair and rank ordering judgements. It is worth noting that the study reported in King *et al* (2009) used alternative research methods to verbal protocols or total mark to measure correlations.

The verbal protocol analysis results indicated that the responses to some items were more influential in contributing to judgements than others. Previous research indicated that in awarding, Thurstone pairs and rank ordering subject experts paid attention to valid features of script quality (Cresswell, 1997; Crisp, 2007; Edwards and Adams, 2003; Greatorex, 2009). Additionally, Greatorex (2009) found that part of the decision making process in all conditions included comparing the answers to one question from the archive examination with answers to a question testing similar knowledge and skills on the live examination. Arguably such a strategy leads to valid

comparisons.

Two items resulted in statistically significant differences between the mean item level marks of candidates who gained a grade A and grade B. These items were thought to provide sound evidence for making decisions about grading standards. The item analysis justified the relatively frequent referencing of item VIII in *rank ordering* and *Thurstone pairs visible*, as well as the relatively frequent referencing of item XII in *awarding clean*. However, generally, the participants did not reference the statistically discriminating items relatively frequently. This reflected the findings of Greatorex *et al* (2008) regarding live scripts. This is to be expected as research literature from overseas indicates that subject experts have varied success in predicting item level statistics including how well different groups of students will perform (Cadwell, 1950; Impara and Plake, 1998; Hambleton and Jirka, 2006). If subject experts have varied success in predicting how well different groups of candidates will perform on an item, they will also have varied success identifying which items discriminate between the achievement of item level marks of grade A and grade B candidates. Perhaps awarders and comparability study participants would benefit from item level analyses to help them identify good items for referencing, although this might be making the judgement process too prescriptive. Of course this would not be possible if rank ordering or Thurstone pairs were to replace marking and grading in one seamless process as suggested by Pollitt and Elliott (2003a and b), Pollitt (2004) and Kimbell (2007).

## *Recommendations for future comparability research*

The verbal protocol analysis indicates that some items contribute more than others to the subject experts' decisions. However, the verbal protocol data and associated decisions are from too few scripts to statistically explore to what extent the item level marks from various items contribute to a Rasch measure of script quality. In future research one way to explore how the construct(s) measured by individual items contribute to the measure of script quality is to correlate item level marks with measure of script quality for all items, and produce associated scatterplots. Such analyses will be facilitated by the routine collection of item level marks in on-screen marking. Such analyses would also help to confirm the construct being measured.

Generally the mark to measure correlation in comparability studies is good. Bramley (2007, 279) says that if the total mark to measure relationship is not good "it casts doubt on the validity of the exercise".

However, if there are strong correlations between some items and the measure of script quality, even when the total mark to measure correlation is lower than hoped there might be some validity in the exercise.

The finding that some items are more important in decision making than others might help explain why we sometimes have the unusual case of rank ordering studies which do not result in robust comparisons. However, to explain the link we must first consider the design of rank ordering studies and associated practices in more detail. Afterwards the connection between my finding and comparability study design will be explored.

For many successful rank ordering studies scripts are assigned to packs which overlap in total marks. The following is an over simplified example but it illustrates the principle behind pack design. A pack might have five archive scripts with total marks 0 to 10 as well as five comparison examination scripts with total marks 0 to 10. The next pack might have five archive scripts with total marks 6 to 20 as well as five comparison examination scripts with total marks 6 to 20. The pattern would continue for the whole mark scale of each examination, so that the total marks in the various packs overlap. This avoids making unnecessary judgements, e.g. comparing scripts with very high and very low total marks, which are almost certainly going to result in the high scoring script being judged to be better. For a detailed discussion see Bramley *et al* (2008). Generally the Bramley *et al* (2008) approach to pack design works well and there is an impressive mark to measure correlation. The question "is the impressive total mark to measure correlations an artefact of pack design?" has been raised. Bramley *et al* (2008) use data and analogies to argue that the answer to this question is "no".

Not all rank ordering studies use this approach to pack design. The major part of the pack design in Curcin *et al* (2009) was generated by random allocation. Raikes *et al* (2009) used some of the usual

principles of pack design, but their study had the novel feature that they also checked that the item level marks of scripts fitted the Rasch model before they included any given script in their sample. Pollitt (2004) and Kimbell *et al* (2007) suggested and used an approach based on the 'Swiss rules' system used in Chess tournaments to determine who plays whom. In their approach experts made decisions in rounds:

Round 1 - scripts were assigned to 'packs' and judges at random.

Round 2 – scripts which were generally ranked high were concentrated in packs, and scripts which were generally ranked as low were concentrated in other packs.

Round 3 – as for round 2. In addition scripts which received a mixed response (ranked both high and low) were included in as many packs as possible.

The Swiss rules approach focuses experts on making judgements where they are most needed for the statistical analysis.

In unusual rank ordering circumstances, problems can arise if there are few judgements (or scripts), particularly if the judgements about scripts linking packs are missing and / or do not follow the expected pattern. In such a situation the data would be disjointed, and there would be no total mark to measure correlation, and robust comparisons cannot be made. One of the findings from my analysis of verbal protocol data was that judgements are weighted towards particular items. These items might or might not be a good proxy for total mark. *If* the latter is the case when subject experts are judging linking scripts they are likely to decide that the scripts with the higher total marks are of worse quality than the scripts with lower total marks. Such decisions might help explain why some rank ordering judgements do not follow the expected pattern, and might contribute to a flawed study.

Perhaps further debate and refinement of methods would be helpful given the variety in approaches to script allocation, as well as the potential link between subject expert's judgement processes and weak total mark to measure correlations (see above). Such research might employ a Latin square design. For instance three groups of subject experts matched for experience in comparability exercises would be recruited. Three mutually exclusive samples of scripts would be created and matched for characteristics like item and total marks, sex of the candidate, centre type and so on. The design ensures that each participating subject expert group would experience the approaches to script allocation in a different order, and each script allocation approach would be used with all three script samples.

# CONCLUSIONS

In summary there are two key findings.

Firstly, subject experts referenced some items relatively more frequently than others and the most referenced items varied with condition. This suggests that slightly different constructs are measured in each condition, which is somewhat at odds with previous research findings. Research is still ongoing to confirm the constructs used to make comparisons in Awarding, Thurstone pairs and rank ordering. Additionally, a way of confirming the constructs used in comparability studies is suggested, as well as a way of researching issues of script allocation.

Secondly, two items statistically discriminated between item level marks of candidates who were awarded grades A and B. These two items were not always the most referenced items. This is in line with previous research about subject experts' predictions about how well students will perform on items. This second finding supports the use of item level data in Awarding to help Awarders focus their judgements on appropriate items.

# REFERENCES

Angoff, W. H. (1971). *Scales, norms, and equivalent scores. In* R. L. Thomdike (Ed.), Educational measurement (2nd ed., pp. 508-600). Washington, DC: American Council on Education.

Arlett, S. J. (2003) *A Comparability study in VCE Health and Social Care, Units 3, 4 and 6: a*

A study based on the Summer 2002 examination and organised by AQA on behalf of the Joint Council for General Qualifications.

Backlund, L., Skånér, Y., Montgomery, H., Bring, J. and Strender, L-E. (2003) Doctors' decision processes in a drug-prescription task: The validity of rating scales and think aloud reports. *Organizational Behaviour and Human Decision Processes*, 91, 1, 108-117.

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

Bramley, T., Gill, T. and Black, B. (2008) *Evaluating the rank-ordering method for standard maintaining*. Paper presented at the International Association for Educational Assessment annual conference, Cambridge. http://www.cambridgeassessment.org.uk/ca/digitalAssets/171223_TB_TG_BB_Evaluating_rankorder_IAEA2008.pdf

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

Cadwell, D. H. B. (1950) Accuracy of prediction of item difficulty for a recent civil service examination for clerks. *Canadian Journal of Psychology*, 4, 1, 18-25.

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, September, Baku, Azerbaijan.

Curcin, M., Black, B. and Bramley, T. (2008) *Standard-maintaining by expert judgement: using the rank-ordering method for determining the pass mark on multiple-choice tests*. A paper presented at the British Educational Research Association Annual Conference, September 2009, Manchester

Cushing Weigle, S. (1999) Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing Writing*, 6, 2, 145-178.

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

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

Ericsson, K. and Simon, H. (1980) Verbal reports as data. *Psychological Review*, 87, 3, 215-251.

Ericsson, K. and Simon, H. (1993) *Protocol analysis: Verbal reports as data.* MIT Press: Cambridge, MA.

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

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

Greatorex, J. (2009) How do examiners make judgements about standards? Some insights from a qualitative analysis. A paper presented at the American Educational Research Association conference, April, San Diego, USA.

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

Greatorex J. and 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 British Educational Research Conference, September 2008, Heriot Watt University, Edinburgh.

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

Green, A. (1998) *Studies in language testing, 5: Verbal protocol analysis in language testing research*. Cambridge University Press: Cambridge.

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.

Hambleton, R. K. and Jirka, S. J. (2006) Anchor-Based methods of judgementally estimating item statistics. *In* S. M. Downing & T. M. Haladyna. *Handbook of test development*, pp 399-420. Lawrence Erlbaum Associates, Inc. New Jersey.

Hambleton, R. K. and Pitoniak, M. J. (2006) Setting performance standards. *In* R L Brennan (Ed.) *Educational Measurement* (4th Edition), pp 433-470. American Council on Education and Praeger Publishers: Westport.

Impara, J. C. and Plake, B. S. (1998) Teachers' ability to estimate item difficulty: a test of the assumptions in the Angoff standard setting method, *Journal of Educational Measurement*, 35, 1, 69-81.

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

King, P., Novakovic, N, and Suto, I. (2009) Capturing expert judgement in grading: an examiner's perspective, *Research Matters A Cambridge Assessment Publication*, 8, 32-33.

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

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

Ofqual (2008) *Making a difference Promoting confidence in A level and GCSE exams in England: summer 2008* Ofqual/08/3984 http://www.ofqual.gov.uk/files/090-Making-a-Difference.pdf

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

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

Pollitt, A. (2004) Let's stop marking exams. Paper presented at the IAEA Conference, Philadelphia, June 2004. http://www.cambridgeassessment.org.uk/ca/digitalAssets/113942_Let_s_Stop_Marking_Exams.pdf

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

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

Suto, W. M. I and Greatorex, J. (2008) What goes through an examiner's mind? Using verbal protocols to gain insights into the GCSE marking process, *British Educational Research Journal*, 34, 2, 213 - 233

Raikes, N., Scorey, S. and Shiell, H. (2009) Grading examinations using expert judgements from a diverse pool of judges, *Research Matters: A Cambridge Assessment Publication*, 7, 4-8.

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: Victoria, Australia.

---

i A *grade boundary* is the lowest mark a candidate needs to be awarded a particular grade.

ii Candidates' performance from the examination or assessment. This is usually a written text. However, it might also be any of the following: a video of the candidate performing a dance or drama, an artefact such as a painting or photograph of something designed and made in a wood work session.

iii Scripts from previous examinations.

iv There are generally three or more units in a qualification, and generally each unit has one question paper, examination or assessment e.g. coursework.

v Operational is used to mean the current year's examination which is not part of an experiment and results in marks and grades on candidates' certificates.

vi Alternatively, Thurstone pairs or rank ordering could be used to compare candidates' performance in cognate qualifications from the same examination session, in which case different script samples would be used and a slightly different statistic would be utilised.

vii *Live* is used here to mean the examination for which the boundary (or boundaries) is to be identified in a study or experiment.

viii In the summer of 2008 Ofqual regulated remote Awards at EdExcel and OCR. Ofqual (2008, 8) concluded that "the remote awarding systems worked well".