or three tied rankings in all 48 rank orders. The contents of each pack were systematically varied in terms of both the overall level and spread of scripts from each year. The judges were warned not to make any assumptions about the contents of their packs – it was possible (for example) for all five scripts from one year to be 'better' than all five scripts from the other.

The data were analysed with two statistical methods, both based on the Rasch model. The first method converted each ranking into a set of paired comparisons and proceeded to analyse them as usual. The second method treated each ranking as a separate Rasch Partial Credit item. When the resulting measures from the two methods were plotted against each other the points lay on a straight line, showing that the two methods were giving substantively the same result.

More interesting was the outcome of the exercise, obtained by plotting the mark on the script against the judged measure, and fitting separate best fit lines for each year, as shown in Figure 1.

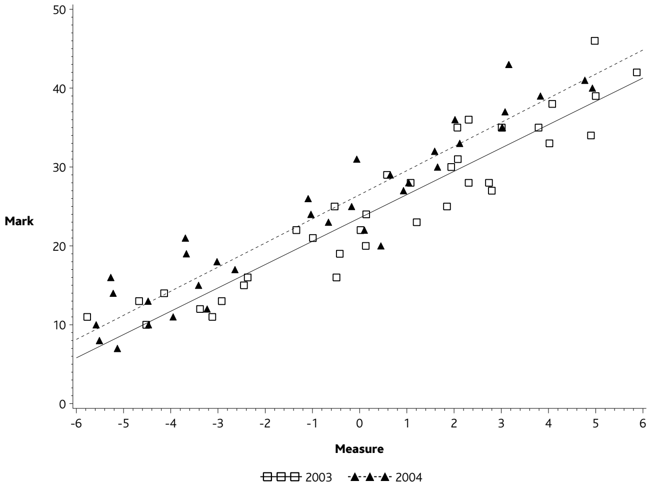

Figure 1: Plot of mark against measure for scripts from 2003 and 2004.

Since the judged measures are all on the same scale, the two raw mark scales can be equated (perhaps a weaker term such as 'linked' is more appropriate): the marks corresponding to the same measure are deemed to be equivalent. The equivalent mark on the 2004 test to any mark on the 2003 test can be found either by reading off the graph, or by using the regression equations for the best fit lines. In fact, in this case the two best fit lines were approximately parallel, separated by a vertical distance of around three marks, leading to the conclusion that the 2004 Reading component was about three marks easier at all levels than the 2003 Reading component. This agreed well with the (completely independent) evidence from statistical equating of pre-test scores, which had suggested that the 2004 test was around two marks easier.

The article contains a lengthy discussion of the difference between standard setting and standard maintaining, arguing that the rank-ordering method is more appropriate than most other judgemental methods for standard maintaining, and that standard maintaining is more appropriate than standard setting in the national testing context.

Since the paper was written, the method has been repeated successfully with the 2004 and 2005 Writing components of the KS3 English test, and is currently being investigated in a research study using scripts from two years of an A-level Psychology paper. There are also plans to investigate its suitability as an award meeting methodology.

Further reading

Bramley, T. (2005). 'A rank-ordering method for equating tests by expert judgement', *Journal of Applied Measurement* **6**, 2, 202–223. Available from http://www.jampress.org

DESCRIBING ACHIEVEMENT

A review of research about writing and using grade descriptors in GCSEs and A levels

Dr Jackie Greatorex Principal Research Officer, Research Programmes Unit

In this article I describe current Awarding practice and review some of the literature about writing and using grade descriptors (often also referred to as 'grade descriptions') for GCSEs and A-levels. Particular emphasis is given to the research that has used empirical evidence to write grade descriptors and the associated research methods.

> *Grade descriptors are descriptions of the qualities expected at different levels of a candidates' performance in an assessment* (Greatorex et al., 2001, 167).

The following are some extracts from grade descriptions for GCSE Biology:

Grade F: Candidates recall a limited range of information. For example, they state the main functions of organs of the human body and describe some defence mechanisms of the body (*OCR, 2000, 17*).

Grade C: Candidates describe how evidence is used to test predictions made from scientific theories, and how different people may have different views on some aspects of science (*OCR, 2000, 18*).

Grade A: Candidates use detailed scientific knowledge and understanding in a range of applications relating to scientific systems or phenomena. For example, they explain how temperature or water content is regulated in humans (*OCR, 2000, 18*).

Awarding

The Code of Practice (QCA, 2005) sets out the procedure by which grade boundaries should be determined. The Awarding Committee (senior examiners) must be provided with a variety of information to set the boundaries, including, where available, performance descriptions or grade descriptions. The Awarding Committee scrutinises scripts within a range of marks around the proposed key grade boundaries (e.g. A/B and E/U for A-level). They start at the top of the range of marks, scrutinising scripts on each mark in turn and agree on the lowest mark that is worthy of a higher grade. This is the upper limiting mark. Then they start at the bottom of the range, scrutinising scripts on each mark in turn and agree on the highest mark that is not worthy of the higher grade. The mark above this is the lower limiting mark. As a group they use their professional judgement to recommend a grade boundary within the range between the higher and lower limiting mark. The grade boundaries for grades which are not key grades are determined by taking the mark interval between key boundaries and dividing it equally between the grades. There are detailed procedures given in the Code of Practice explaining the rules to apply when the mark interval cannot be divided equally between the grades.

Writing grade descriptors

There have been attempts to write prescriptions of candidates' performances to be associated with different grades (grade related criteria) for GCSEs. For further information regarding the development of the grade related criteria and the associated limitations see Gipps (1990), Kingdon and Stobart (1988) and Cresswell (1987). The difference between grade descriptors and grade criteria is an important distinction to make. 'Grade criteria' are qualities a candidate's work *must* exhibit to be awarded a grade. 'Grade descriptors' are indicators which exemplify the qualities candidates are likely to exhibit if they achieve a particular grade. Normally grade descriptors refer to mid range performance within a grade rather than borderline performance. This article focuses on grade descriptors.

Massey (1982) attempted to use empirical evidence to describe the performance of candidates who achieved particular grades. He analysed marks at the question level, adopting a concept of group mastery as the aim was to describe the achievement of grade groups, not individuals. A grade group is all the candidates who were awarded a particular grade. If the mean mark achieved by a grade group was 75% of the total marks available for a question this was taken as an indication of group mastery. The arbitrary working value of 75% suggests that most of the candidates can answer a question correctly. There were two criteria for mastery:

- questions had to be mastered by all grade groups higher than the mastery level grade group and all groups below had to fail to reach the 75% criterion;
- questions had to discriminate statistically between the mastery level grade group and the group below (Massey, 1982).

The skills and knowledge required to answer questions where a given grade group defined the mastery level can be used to describe the grade group's competency. For the lower ability range the analysis did not indicate what candidates could master because on these tests the lower ability candidates generally got less than 75% on all questions. Massey recommended that other approaches would need to be used to capture this information. He argued that if other educators produced written descriptions of performance at the grades featured in his study, their descriptions would need to be reconciled with his results, or they would lack validity.

Pollitt and Murray (1996) argued that grade descriptors should match what assessors perceive in the performance they assess. Their method for writing grade descriptors was designed with this in mind. They asked assessors to compare pairs of candidates' work in the field of testing English as a second language. In each comparison they indicated which performance was better and which was weaker. The judgements of which performances were the better in each pair were statistically analysed to create a scale on which the performance of the candidates could be located. Immediately after making a judgement the assessors were interviewed using Kelly's Repertory Grid (Kelly, 1955) to describe the two performances. Kelly's Repertory Grid is a method of interviewing research participants in a systematic way to compare how objects – in this case candidates' work – are similar to and different from one another. It is a way of eliciting peoples' personal constructs. They found that particular characteristics of performance seem to be allied to different sections of the scale. The characteristics of low performance became increasingly less relevant at the higher performance end of the scale, while different characteristics exemplified higher performance and were evident only at the higher end of the scale. Pollitt and Murray argued that assessment judgements would be more accurate if grade descriptors referred only to the characteristics which are normally salient at each stage on the performance scale.

The following section refers to three articles about developing grade descriptors for A-levels and a fourth article about their use in teaching.

(1) Making the grade – Developing grade descriptors for Accounting using a discriminator model of performance, Greatorex, J., Johnson, C. and Frame, K. (2001), *Westminster Studies in Education*, **24**, 2, 167–181.

One of the purposes of the research was to establish whether the writing of grade descriptors for an Accounting A-level, examined in June 1998, was aided by the discriminator model of performance. The *discriminator model of performance* is a term we used to refer to Pollitt and Murray's argument that candidates exhibit distinctive qualities at different stages on a performance scale. Pollitt and Murray do not use this term themselves. The methods we used draw from both Massey, and Pollitt and Murray.

We took a random sample of candidates who achieved each component (question paper) grade A, B, C, D, E and O/N. The random sample for each grade is assumed to represent candidates' achievement at that grade on each question in the examination paper. A series of statistical criteria for mastery, similar to those used by Massey, were applied to these data. This stage of the research is called a *mastery levels analysis*. The outcome of the analysis is a list of examination questions which particular component grade groups have 'mastered'. On these questions candidates from adjacent grade groups are exhibiting different knowledge or skills.

The second stage of the research was to describe the qualitatively different knowledge and skills exhibited by candidates from different component grade groups. Two senior Accounting examiners were presented with two answers to a question from candidates in the grade group which mastered the question and one script from a candidate from the grade group below. These answers had been credited with the mean number of marks for their grade group on that question. The answers were then compared using an adapted version of Kelly's Repertory Grid. The following extract is an example of which candidates' answers were compared:

> *two E grade scripts where the candidates had scored 4 marks on question 4a on component 3 and one grade O/N script where the candidate had scored 2 marks on question 4a*.
> (Greatorex et al., 2001, 175)

The examiners were interviewed and asked to describe how the answers to each discriminating question on the higher grade scripts were similar to one another and different from the answer on the lower grade script. The interviews were recorded. The procedure was repeated using a small number of scripts for each grade. This provided a record of the knowledge and skills which distinguished the performance of candidates who achieved particular grades from the performance of candidates awarded the grade below. From this record the researchers and senior examiners wrote grade descriptors.

Extracts from the grade descriptors are given below:

> ***Grade E:*** *Knowledge of what is required, understanding, selecting appropriate knowledge, some application, but can be an incomplete answer*.
>
> ***Grade B:*** *Can rise to the challenge of a novel situation, but with some evidence of technical and calculation errors* .
>
> ***Grade A:*** *Awareness of multi-faceted aspects of Accounting* (Greatorex et al., 2001, 176–177).

Writing grade descriptors grounded in empirical evidence is arguably an improvement upon methods which are based upon examiners' expectations alone. However, one problem of using empirical evidence in this way is that it is a post hoc method with grade descriptors based on what candidates achieved rather than on what the examinations were designed to assess.

The grade descriptors developed in this study were validated by using them in the Awarding Meeting for the Accounting A-level examined in June 1999. They were generally received positively because they proved to be appropriate for the 1999 scripts. This implies that the discriminator model of performance is a sound basis for a method of developing grade descriptors in this domain. The grade descriptors which were developed using the discriminator model of performance suggested that there are indeed different characteristics associated with each grade. However, in some cases the differences might relate to 'relative' performance. The following extract gives an example of relative performance:

> ***Grade A:*** *Clarity and brevity of expression in a focused answer* (Greatorex et al., 2001, 176).
>
> ***Grade B:*** *Focused answer* (Greatorex et al., 2001, 176).

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

In the article referenced above I argued that the process of writing grade descriptors is a way of making senior Accounting examiners' tacit knowledge, or personal constructs about achievement at different grades, more explicit. Other sources of this information can be found in the specifications, question papers and mark schemes. Moreover, it is fair to share this knowledge so that teachers and candidates are aware of the qualities expected to be credited with a particular grade. The method described in this article differs from that used in the previous article in that the subject officer and researcher, rather than the examiners, collated the records from the interviews to write grade descriptors. It can be argued that the grade descriptors would be more valid if they were formulated from the interview records by the examiners themselves. The usefulness of the grade descriptors at Awarding was limited, since the style of the question papers had changed between data collection and validation. Nevertheless, they did provide a helpful reference point.

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

In this study I aimed to develop grade descriptors for an A-level in Economics based on examination data and scripts from the summer of 1999. The Economics A-level had three examination papers: a multiple choice paper, a paper with a points based mark scheme and a third paper where the mark scheme was based around four generic level descriptors. Grade descriptors were not developed for the multiple choice paper but were developed for the other item types using the methods utilised in the two studies to write grade descriptors for Accounting described previously. In contrast to the studies in Accounting, the Economics grade descriptors were validated by using them in two Awarding Meetings in the next session, one was the same Economics A-level specification used for data gathering and the other was a different Economics A-level.

The Accounting papers had contained numerical questions as well as long and short written answer questions and the Economics questions were short answers and essay questions. I argued that the method of mastery levels analysis and Kelly's Repertory Grid technique can be used together for examinations with all of these types of questions and for subject domains which are orientated towards both numerical work and extended prose. One advantage of the grade descriptors developed in these studies is that they are written at the component level (i.e. the level at which the judgements are made in Awarding Meetings) which is different from the general practice of writing grade descriptors at the specification level.

Developing grade descriptors from empirical evidence using sound methods is good practice. It is also important that the grade descriptors are used appropriately and in the following section I refer to an article which addresses the issue of using grade descriptors in teaching.

(4) Can different teaching strategies or methods of preparing pupils lead to greater improvements from GCSE to A-level performance?
Greatorex J. and Malacova E. (in press) *Research Papers in Education*.

In this study the aim was to investigate whether there was any relationship between relative progress from mean GCSE scores to A-level results on the one hand and teaching strategies or how teachers prepared pupils for assessments on the other. We sent a questionnaire to Chemistry teachers to survey the teaching strategies they used and how they prepared pupils for the different A2 units which are part of the A-level. The questionnaire responses were matched to A-level and GCSE results. A series of multilevel models were fitted to the data to identify any relationship between relative progress from mean GCSE scores to A-level results and the questionnaire responses. We found some activities which related to higher relative progress from mean GCSE to A-level unit marks. One of these was using grade descriptors to inform the teacher's preparation of pupils for the synoptic unit. However, it was also found that there was no such effect for the other two A2 units, one of which was coursework.

The Chemistry grade descriptors had been developed using examiners' expectations alone rather than based on empirical evidence. Nevertheless, our findings showed that grade descriptors can be important and helpful to teachers and can enhance classroom practice.

Conclusions

It can be deduced that, when resources allow, it is good practice to write grade descriptors based on empirical evidence. It seems that grade descriptors for different domains and types of questions can be written using a combination of a mastery levels analysis and Kelly's Repertory Grid technique. The grade descriptors developed using these methods describe the distinctive characteristics of achievement at particular grades.

Despite the difficulties of effectively communicating the meaning of grade descriptors to examiners, teachers, candidates and other stakeholders, it is good practice to make efforts in this area.

There is little research about how grade descriptors are used, or could be used, in relation to teaching GCSEs or A-levels, or in preparing pupils for assessments and there is room for further research in this area.

References

Cresswell, M. J. (1987). 'Describing examination performance: grade criteria in public examinations', *Educational Studies*, **13**, 3, 247–265.

Gipps, C. (1990). *Assessment: a Teacher's Guide to the issues*. London: Hodder and Stoughton.

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

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

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., Johnson, C. and Frame, K. (2001). 'Making the grade – Developing grade descriptors for Accounting using a discriminator model of performance', *Westminster Studies in Education*, **24**, 2, 167–181.

Greatorex J. and Malacova E. (in press). 'Can different teaching strategies or methods of preparing pupils lead to greater improvements from GCSE to A-level performance?', *Research Papers in Education*.

Kelly, G. (1955). *The psychology of personal constructs*. New York: Norton. Reprinted by Routledge (London), 1991.

Kingdon, M. and Stobart, G. (1988). *GCSE Examined*. London: The Falmer Press.

Massey, A. J. (1982). 'Assessing 16+ Chemistry: The exposure-mastery gap', *Education in Chemistry*, September, 143–145.

Oxford Cambridge and RSA Examinations (2000). OCR GCSE in Biology 1980, www.ocr.org,uk

Pollitt, A., and Murray, N. L. (1996). 'What raters really pay attention to', in M. Milanovic, & N. Saville (Eds), *Studies in Language Testing: 3 Performance Testing, Cognition and Assessment: selected papers from the 15th Language Testing Research Colloquium*. Cambridge: Cambridge University Press.

Qualifications and Curriculum Authority (2005). *GCSE, GCE, VCE, GNVQ and AEA Code of Practice 2005/6*. London: QCA.

ISSUES IN QUESTION WRITING

Can a picture ruin a thousand words? The effects of visual resources in examination questions

Victoria Crisp and Ezekiel Sweiry Research Officers, Research Programmes Unit

Introduction

Visual resources, such as pictures, diagrams and photographs, can sometimes influence students' understanding of an examination question and their responses (Fisher-Hoch, Hughes and Bramley, 1997). Visual resources are sometimes included to test students' abilities to interpret them, but they are more commonplace than this alone would warrant.

Research on the influences of graphics in instructional texts provides some relevant insights. Such research has often found illustrations to have a positive influence on learning and retention (Weidenmann, 1989; Ollerenshaw, Aidman, and Kidd, 1997). However, the main purpose of examination questions is to assess learning rather than teach. Graphics are thought to 'simplify the complex' and 'make the abstract more concrete' (Winn, 1989, p. 127). Graphics can also provide more information than can be explained in words (e.g. Stewart, Van Kirk and Rowell, 1979). These are justifiable reasons for including visual resources in examinations as they can reduce the length of questions and help students to access abstract concepts. In addition, illustrations are generally believed to have a motivational role in the context of instructional texts (Peeck, 1993) which could apply to examinations.

In their review of work in this area Levie and Lentz (1982) found that in about 15% of studies there were no significant effects of including illustrations. One possible explanation is that the quality and appropriateness of the graphic is important (see Peeck, 1987 for some evidence of this). Such failures have also been explained as either a result of students' learning styles (as Ollerenshaw, Aidman and Kidd, 1997 report) or due to students not processing graphics adequately (Weidenmann, 1989). The latter is thought to be a result of the apparent ease of processing an illustration, giving students the false impression