'Tools for the trade': What makes GCSE marking reliable?

A paper presented at the conference *Learning Communities and Assessment Cultures: Connecting Research with Practice.* The conference was jointly organised by the EARLI Special Interest Group on Assessment and Evaluation and the University of Northumbria.

28-30 August 2002, University of Northumbria UK.

Jackie Greatorex, University of Cambridge Local Examinations Syndicate

Jo-Anne Baird, Assessment and Qualifications Alliance

John F. Bell, University of Cambridge Local Examinations Syndicate

Disclaimer

The opinions expressed in this paper are those of the authors and are not to be taken as the opinions of the University of Cambridge Local Examinations Syndicate (UCLES) or any of its subsidiaries.

Note

This research is part of a programme of collaborative work undertaken by the Assessment and Qualifications Alliance (AQA) and the University of Cambridge Local Examinations Syndicate (UCLES) about the consistency of marking. The participants in the research were examiners who mark for AQA and Oxford Cambridge and RSA Examinations (OCR). OCR is a subsidiary of UCLES.

Contact details

Jackie Greatorex

Research and Evaluation Division, University of Cambridge Local Examinations Syndicate, 1 Hills Road, Cambridge, CB1 2EU.

☎ 01223 553835

FAX: 01223 552700

greatorex.j@ucles.org.uk

This paper is available at www.ucles-red.cam.ac.uk

'Tools for the trade': What makes GCSE marking reliable?

Abstract

The Qualifications and Curriculum Authority requires Awarding Bodies to use standardisation procedures (e.g. co-ordination meetings and providing exemplar scripts marked by the Principal Examiner (PE)) to make GCSE marking reliable. AQA and UCLES collaboratively researched these procedures. In the first experiment with English examiners, it was found that if they were given exemplar scripts at the centre of the marks associated with a level descriptor rather than exemplar scripts at the lowest marks for the level their marking was more severe. In the second experiment, with History examiners it was found that different styles of co-ordination meetings and whether a co-ordination meeting was held did not affect the inter-rater reliability of the marking.

This paper reports on a survey of History examiners, which was undertaken after the second experiment. Questionnaires about aspects of standardisation were distributed to the examiners who took part in the experiment and 35 examiners responded. The data were quantitatively and qualitatively analysed.

The survey shows that all aspects of standardisation are important, particularly the mark scheme and the co-ordination meeting. This is in contrast to the experimental results (above). Examiners value the opportunity in the co-ordination meeting to develop a 'community of assessment practice' and learn about the application of the mark scheme. The co-ordination meeting is useful as it gives examiners a feeling of being part of a team, boosts confidence and provides examiners with feedback.

The literature review suggests that exemplar scripts and discussion between examiners is important in facilitating reliability. In the survey of History examiners reported here exemplar scripts and discussion between examiners were rated as useful as they facilitated the understanding and application of the mark scheme. The literature says that not everything can be written down and that some understanding of how to apply the mark scheme will remain tacit. But the History examiners surveyed here pointed out that the mark scheme and how it is written are important. They thought that exemplar scripts should be annotated and related to the levels in the mark scheme. The experiment in English illustrated that the way that the exemplar scripts are tied to the mark scheme affects the severity of marking.

A principle of a community of practice is that practice is negotiated in a non-hierarchical manner so that there is shared understanding and ownership. Some examiners found hierarchical discussion useful, but others found non-hierarchical discussion useful.

Introduction

In this paper two experiments about the training of GCSE examiners are described, but the main focus is on a survey of examiners who participated in one of the experiments. The General Certificate of Secondary Education (GCSE) is a qualification taken by sixteen year olds in England. GCSEs are offered in a wide variety of subjects and are administered by three Awarding Bodies; the Assessment and Qualifications Alliance (AQA), the EdExcel Foundation and Oxford, Cambridge and RSA Examinations (OCR); OCR is a subsidiary of the University of Cambridge Local Examinations Syndicate (UCLES). One of the responsibilities of Awarding Bodies is to ensure that the marking of their assessments is reliable, and the research reported here is part of a collaborative program of research undertaken by AQA and the Research and Evaluation Division of UCLES into ways to improve the reliability of marking in GCSE examinations. Standardisation is one of the processes used to make the marking more reliable: it involves the use of an agreed mark scheme, a co-ordination meeting to brief the examiners about the use of the mark scheme and monitoring of examiners' marking. The procedures used throughout the examination process are detailed in the Code of Practice (Qualifications and Curriculum Authority, 2002). The Code of Practice is written and published by the Qualifications and Curriculum Authority (QCA), the government body which regulates the Awarding Bodies. The research draws from a theory of 'communities of practice' and research literature about the reliability of marking.

The AQA and UCLES work is not the only recent work in assessment to draw from the theory of communities of practice. Konrad (1998) argued that the reliability of assessment in vocational qualifications in the UK would be improved by the introduction of a community of practice. At the same time Wiliam (1998, 9) argued that:
This notion of "understanding the standard" is the theme that unifies summative and formative functions of assessment. Summative requires that teachers (or other assessors) become members of a community of practice, while formative assessment requires that the learners become members of the ***same*** *community of practice.*

In this context he used the word 'standard' to refer to standards for individual tasks (say an examination question) as well as overall assessments (for example, an examination). Later Hall and Harding (2002) coined the phrase a 'community of assessment practice' in their investigation of whether communities of assessment practice existed in UK primary schools for the purposes of facilitating the consistent application of assessment criteria from the National Curriculum in English.

Researchers have found a multitude of factors that affect the reliability of marking. For example, different types of marking can lead to different levels of reliability e.g. Hartog and Rhodes (1935) and later Britton et al. (1966). Lenney et al. (1983) found that making the mark scheme more specific increases reliability. In the case of English as a foreign language (EFL), Weigle (1998), Stahl and Lunz (1991) and Lunz et al. (1990) found that training could improve the consistency of each individual examiner's marking (intra-rater reliability). However, Lunz and O'Neill (1997) found that retraining does not affect the leniency and severity of examiners. More recently, Shaw (2002) found that an iterative standardisation process of training and feedback to EFL examiners did not improve inter-rater reliability (consistency between examiners) but that inter-rater reliability was consistently high any way. He argued that the mark scheme itself has a strong standardising effect. He also added that examiners did modify their behaviour after each round of training. Also in the area of EFL, Wigglesworth (1993) experimented with providing feedback to EFL examiners as part of a training and standardisation process. She found that examiner consistency improved and that biases reduced following feedback.

Wolf (1995) argued that in assessment systems the use of examples of candidates' work was particularly important as the standard is illustrated by the candidates' work rather than by descriptions of their work. She added that there is little research into the importance and role of examples of candidates' work in achieving consistency of judgement. With these points in mind it was decided that a study (study 1) should investigate the use of exemplars in the standardisation of marking in GCSE examinations which use a banded mark schemes (Baird et al., 2002).

A banded mark scheme is one which has a series of descriptors each associated with a band of marks. When examiners use the mark schemes they apply a principle of best fit. For example, when an examiner reads a candidates' answer they decide which level descriptor best describes the answer and then choose an appropriate mark from the range available in that band. It should be stressed that the level descriptors are indicators not criteria and that they are compatible with a principle of compensation within the mark scheme, that is candidates can compensate for their weaknesses by gaining extra marks for their strengths.

Wolf (1995) also argued that discussion in tight assessor networks facilitates reliability and that standards are communicated by examples of students' work rather than by written criteria or indicators. Wolf's recommendations fit well with the notion of communities of practice, which are tight networks or teams in which people can learn from one another yet maintain a shared ownership of the social practice. Additionally, it is argued that learning is facilitated by a "flat hierarchy" (Wenger, 1998). A flat hierarchy would mean that members of a community might have different roles and responsibilities and different levels of authority but decision making would be a shared experience.

For a university Communication and Media examination, Barrett (2000) investigated the inter-rater reliability, examiner leniency and other types of errors like the halo effect. He found that one examiner was particularly free of errors, and argued that this was an issue of ownership as the error-free examiner had set the test and was a senior lecturer for the course associated with the examination. It seems that true ownership of a marking scheme will lead to accurate marking, and suggests that a feeling of joint ownership amongst a team of markers might lead to more reliable marking by all. Given that flat hierarchies, discussion and shared ownership are likely to facilitate consistency of judgement it was decided that an experiment (study 2) should be undertaken to identify whether different types of co-ordination meeting (hierarchical discussion and non-hierarchical discussion) would affect the consistency of marking (Baird et al., 2002).

In study 1, English GCSE examiners for one paper were given exemplar scripts marked by the Principal Examiner (PE) with feedback about why those marks were awarded. One group received no exemplar scripts, another received exemplar scripts at the centre of the marks associated with a level descriptor (prototypical scripts) and another received exemplar scripts at the lowest marks for the level (threshold scripts). Whether examiners did or did not use exemplar scripts and the type of exemplar script did not affect the accuracy of marking, that is the ***absolute differences*** between markers' marks and the PEs marks were not statistically significantly different. However, when the ***actual differences*** were analysed it was found that the group who had prototypical scripts were stricter than the other groups (Baird et al. 2002). However, the results of the studies 1 and 2 must be interpreted with the caveat that these results were raw marks from experimental marking to which additional quality procedures, like scaling and the monitoring of examiners' marking had not been applied.

In study 2, GCSE History examiners took part in the marking of a small sample of scripts. They were then divided into three groups; two of these groups attended co-ordination meetings. The first group attended a hierarchically organised co-ordination meeting where the Principal Examiner decided how the mark scheme should be interpreted, a second group attended a more consensual co-ordination meeting that involved discussion and consensual decision making, and a third group that did not attend a co-ordination meeting at all. After the meetings, the examiners marked a second sample of scripts. In the event there were no statistically significant differences between the inter-rater reliability of the groups. These results might be explained as follows. The markers had marked History paper 1 in the summer 2001 session but in this study they marked paper 2. Nevertheless the community of assessment practice from the live paper 1, the community of practice amongst teachers teaching History at GCSE, and the paper 2 mark scheme were strong enough to facilitate reliability in the experimental marking of paper 2 (Baird et al., 2002). This accords with Furneaux and Rignall's (2000) finding that the mark scheme for an EFL test had a standardisation effect even without the examiners being trained in its use. Alternatively, it might be that these examination questions elicited responses that were unusually close in nature to the level descriptors and therefore comparatively easy to relate to the levels in the mark scheme.

The survey

The examiners from study 2 were also sent a questionnaire about co-ordination meetings, which gave insights into aspects of a community of assessment practice (see appendix A). The questionnaire specifically referred to the issues that have been investigated in the experiments, i.e. mark schemes, co-ordination meetings, exemplar scripts and discussion with other examiners.

The questionnaire was administered to all the examiners who had taken part in the experiment. The questionnaires were sent to the examiners after they had returned their final batch of marking. Forty-five Assistant Examiners were recruited to take part in the experiment and 35 of them returned a completed questionnaire.

For questions 1i, 1ii, 1iii, 1iv and 2 descriptive statistics of the ratings from all respondents were calculated. The ratings that all the examiners made on 1i, 1ii, 1iii and 1iv were explored using correlations. The responses of the examiners in the different experimental groups were also scrutinised for questions 1i, 1ii, 1iii and 1iv by using independent t tests. The free text responses were coded and summarised question by question. They were also coded and subjected to principal components analysis, but it showed no strong pattern to the examiners' views, so the results have not been given here.

Results from the rating scales (questions 1i, 1ii, 1iii and 1iv)

Frequency tables and descriptive statistics for the four parts of question 1 are given below. This question asked the examiners how useful mark schemes, co-ordination meetings, discussions with other examiners, and exemplar scripts were for standardising marking. They used a seven-point scale with 7 indicating 'useful' and 1 indicating ' not useful'.

Table 1 Summary statistics for rating scales

Question	Mean	Standard Deviation	Minimum	Maximum
1i - Mark Scheme	6.25	1.12	3	7
1ii - Co-ordination meeting	6.25	1.06	3	7
1iii - Discussion with other examiners	5.17	1.80	1	7
1iv - Exemplar scripts	5.40	1.26	3	7

On average, the mark scheme and co-ordination meetings were considered to be as useful as one another and were more useful than exemplar scripts. Discussion with other examiners was considered to be the least useful of the four parts of standardisation. However, the ratings were all above the midpoint indicating that each feature of the standardisation process was generally considered to be useful by a majority of the examiners.

Table 2 Correlations between the ratings examiners made on each part of question 1

Pair	Correlation	Significance of the correlation	Significance of t test (2-tailed)
1i versus 1ii	0.238	0.168	1.000
1i versus 1iii	-0.153	0.380	0.008
1i versus 1iv	-0.237	0.171	0.002
1ii versus 1iii	0.144	0.408	0.002
1ii versus 1iv	0.140	0.066	0.001
1iii versus 1iv	-0.237	0.171	0.583

The low correlations in Table 2 indicate that the examiners did not rank the mark schemes, co-ordination meeting, discussion with other examiners and exemplar scripts in the same order as one another. This is hidden by the descriptive statistics in Table 1. There are significant differences between the means i.e. significant differences for the t tests at the $p>0.05$ and $p>0.01$ levels for 1i versus 1iii, 1i versus 1iv, 1ii versus 1iii and 1ii versus 1iv. This indicates that:-

- the mark scheme was judged to be significantly more useful than discussion with other examiners;
- the mark scheme was judged to be significantly more useful than the exemplar scripts and associated marks for the candidates;
- the co-ordination meeting was considered to be significantly more useful than discussion with other examiners;
- the co-ordination meeting was considered to be significantly more useful than the exemplar scripts and associated candidates' marks.

However it is obviously difficult to consider the exemplars, mark scheme, co-ordination meeting and discussion with other examiners as separate entities.

To compare the ratings of the different experimental groups on each question ANOVA was used. No statistically significant differences were found. So the different experimental conditions did not effect how useful the examiners found the mark scheme, the co-ordination meeting, discussion with other examiners and the exemplar scripts with associated marks awarded by the Principal Examiner. Independent samples t tests were also used to investigate any differences between the views of the different aspects of the standardisation procedures. Table 3 generally illustrates that there is little difference between the experimental groups in how they perceived the usefulness of the co-ordination meetings, mark scheme, discussion between examiners and exemplar scripts. However, the number of examiners in some of the groups are small and therefore it is difficult to make generalisations.

The only statistically significant difference between the means indicated by the independent samples t tests is that the control group considered discussion to be more useful than the examiners who attended a hierarchical co-ordination meeting. It seems that when examiners are deprived of the opportunity to discuss the candidates' work in relation to the mark scheme they feel that it would be useful but when they have a discussion in a hierarchically organised co-ordination meeting it was considered to be less useful.

Table 3 Independent samples t tests

Question	Experimental Group	Mean	Standard deviation
1i	control	6.167	1.200
1i	hierarchical	6.222	1.302
1i	Non-hierarchical	6.500	0.756
1ii	control	6.611	0.698
1ii	hierarchical	5.889	1.269
1ii	Non-hierarchical	5.875	1.356
1iii	control	5.667	1.372
1iii	hierarchical	4.000	1.803
1iii	Non-hierarchical	5.375	2.264
1iv	control	5.389	1.335
1iv	hierarchical	5.000	1.225
1iv	Non-hierarchical	5.875	1.126

Question	Experimental Groups compared	Significance level (2-tailed)
1i	control versus hierarchical	0.913
1i	control versus non-hierarchical	0.479
1i	hierarchical versus non-hierarchical	0.605
1ii	control versus hierarchical	0.141
1ii	control versus non-hierarchical	0.077
1ii	hierarchical versus non-hierarchical	0.983
1iii	control versus hierarchical	0.013
1iii	control versus non-hierarchical	0.687
1iii	hierarchical versus non-hierarchical	0.184
1iv	control versus hierarchical	0.417
1iv	control versus non-hierarchical	0.379
1iv	hierarchical versus non-hierarchical	0.148

Results from the free text responses (questions 1i, 1ii, 1iii and 1iv)

The following is a summary of the points made by the examiners in response to questions 1i, 1ii, 1iii and 1iv. For many of the points that were made there were low numbers of examiners making that point and so the following must be seen as a complete description of the range of examiners' views.

1. How useful do you think each of the following are for standardising marking?

i) *Mark scheme*

The mark scheme was described as 'essential' by twenty-two examiners, one examiner described it as *'the tools for the trade'*. Two examiners described the mark scheme as 'prescriptive' and eighteen examiners said that that the mark schemes gives the marking criteria and acts as a guideline to the number of marks to award. Two examiners pointed out the limitation of mark schemes - that not everything can be written down and some 'jargon' or 'short hand' in the mark scheme needs to be explained at a co-ordination meeting. One examiner noted that *'the precise definition of what is required does not necessarily become clear until this has been discussed at a co-ordination meeting e.g. what exactly constitutes explanation rather than description?'*. Thirteen examiners pointed out that the mark scheme must be clear and contain examples of potential candidates' answers. Six examiners commented that they favoured levels of response mark schemes. One commented that *'Mark schemes which operate by level marking are particularly good'*. Five examiners noted that they personalise the mark scheme for their own use e.g. by making annotations. Five pointed out that the mark scheme is a starting point, which needs to be used alongside other forms of standardisation.

ii) *Co-ordination meeting*

The co-ordination meeting was also considered to be 'essential' by eighteen examiners. It was recognised that the co-ordination meeting clarifies queries about what is and isn't allowed in terms of idiosyncratic answers (sixteen examiners). Ten examiners noted that example scripts and answers were marked and discussed to help examiners understand the mark scheme. For instance one examiner commented that *'my ideas are reinforced by exchange of views within a team over a question/answer'*. Additionally nine examiners explained that the co-ordination meeting makes the PE's interpretation of the mark scheme clear in relation to the scripts, for instance, *'useful as we could see how the PE interpreted the mark scheme'*. The co-ordination meeting was considered by four examiners to be a way of boosting examiner confidence *'I felt I was marking more consistently and with more confidence after the meeting'*. According to three examiners the co-ordination meeting offers an opportunity for the mark scheme to be adjusted. However three examiners recognised that it does leave room for examiners to interpret the mark scheme in different ways. It was recognised that the co-ordination meeting is useful because it comes at a time when examiners are learning how to apply the mark scheme through marking their first set of scripts (three examiners). One examiner who was in the control group and experienced no co-ordination meeting in the experiment was unhappy without it and felt that they had missed it. It was also mentioned by one examiner that the meetings avoid examiners being isolated and a member of the control group said *'No sense of being part of a team in the research marking process'*. There are also some negative aspects or limitations of co-ordination meetings. For instance, five examiners said that co-ordination meetings can be too long and include some time wasting.

iii) Discussion with other examiners

Eight examiners considered discussion to be essential: two reported that discussion with the PE was useful, and six that discussion with the Team Leader (TL) was helpful. One examiner from the non-hierarchical co-ordination group said that *'The PE helped to change my 'mind-set' a little; much of the time we don't even realise we have one.'*. Additionally eight said that discussion with other Assistant Examiners (AEs) was useful. For example, one examiner said that feedback from discussion was useful: *'it can be reassuring to know that you give the same mark as a colleague'*. But another examiner warned that *'There is always the danger of unwittingly departing from the agreed mark scheme following, say, late night exchange of views with another examiner. Team leaders should be consulted in the first instance for clarification'*. It was felt by five examiners that, like co-

ordination meetings, discussion between examiners could give a feeling of comfort or confidence *'There's comfort in peer support, too'*. It was again mentioned by one examiner that AEs were not necessarily listened to and the mark scheme was already determined by this stage: *'Unfortunately objections by ordinary examiners at a co-ordination meeting are of no use. The mark scheme is a fait accompli'*. But it was also mentioned by two examiners that a range of views or interpretations of the mark scheme could emerge in discussion *'It is also a chance to appreciate the possible range of reactions and interpretations'*.

iv) Exemplar scripts and associated marks awarded by the PE

Sixteen examiners were of the view that exemplar scripts and marks were useful in a co-ordination meeting as they served to confirm the way that the mark scheme should be applied. That is, how the candidates' answers in scripts are linked to the mark scheme levels can be usefully explained in a co-ordination meeting. One examiner summed this up by saying: *'Useful to some degree but without the discussion and explanation at a co-ordination meeting, it can be perplexing to understand how the final marks were given. Best when part of the co-ordination meeting and discussion, where decisions can be explained, rather than on their own'*. According to eighteen examiners exemplar scripts can be 'helpful' and 'useful', they were also described as a 'guide' (by two) or as 'clarifying issues' (by five). It was commented by two examiners that they could illustrate how marks and levels are allocated and how to record the levels. They were seen by six examiners as a good starting point, which could be useful at the stage of marking the standardisation sample. Five examiners were of the view that *'Exemplar scripts are useful because they give examples from a range of abilities …'* and two examiners mentioned that they *'are useful because they give guidance on when progress between levels is established…'*. Five examiners suggested that a rationale for why particular levels were awarded was useful: *'These were very useful in understanding how marks had been awarded - though I felt some disagreement I was able to see why the marks had been allocated'*. Three others mentioned that scripts should be annotated: *'What would be better would be for the scripts to be annotated so I could tell when and where and why marks were awarded'*. Two examiners agreed that, in the words of one of them: *'Examples attached to mark scheme levels are probably more useful'* than exemplar scripts and associated marks awarded by the PE.

Results from the dichotomous rating question 2

2 Do you think the marking standardisation process was better for the operational examination? (delete as appropriate)Yes/No

If so, why

Table 4 Descriptive statistics

Response	Frequency	Percent	Cumulative Percent
0 (NO)	9	25.7	25.7
1 (YES)	15	42.9	100.0
Total Responses	24	68.6	
Missing	11	31.4	
Total	35	100.0	

Frequency	Valid responses	24
	Missing	11
Mean		0.625
Standard deviation		0.495
Minimum		0.00
Maximum		1.00

There were eleven examiners who did not answer this part of the question. Most of the examiners who answered the question thought that the standardisation process was better for the operational examination. The number of examiners in each group who responded to question 2 was too small to make explorations useful.

Results from the free text responses (question 2)

The responses to question 2 were as follows: five examiners (two from the hierarchical group, one from the non-hierarchical group and two from the control group) wrote on the questionnaire that they did not understand the question. Despite the confusion there were a range of views expressed about the differences between the operational approaches to standardisation and the research project. Eleven examiners, some of whom answered 'yes' and others who answered 'no' to the first part of the question said that they missed having a co-ordination meeting where the ambiguities of the mark scheme are clarified. These examiners were from the control group who did not have a co-ordination meeting so that the effect of co-ordination meetings as well as styles of co-ordination meetings could be identified. Additionally there were not enough examiners in the experiment to warrant a Team Leader structure and five examiners (one in the hierarchical group and 4 in the control group) said that the operational approach was better and that they missed Team Leaders whom they can generally ring for help solving problems. For example, '*Found it very isolating experience, had no-one to phone and ask for clarification or to discuss issues with - Don't think I'd mark if this approach is adopted*'.

Three examiners (two from the hierarchical group and one from the non-hierarchical group) said that there was little difference between the operational and research approaches. This is surprising given that the non-hierarchical co-ordination meeting was designed to be more inclusive. One examiner from the control group said *'about the same - the big difference of not having a co-ordination meeting left a lot of questions unanswered'*. Another three examiners who preferred the operational standardisation approach cited differences between paper 1 and paper 2 as a reason for their preference, which was mostly concerned with familiarity with the paper 2 questions and mark scheme. Two examiners preferred the operational approach because they had access to the PE, which they were denied as a member of the control group. Two examiners (one from the non-hierarchical group and another from the control group) said that they did not prefer the operational process but that feedback was needed for examiners in the research project, for example: *'Feedback on these 10 scripts would have been instructive to us just as it is after the co-ordination meeting in the processes we are used to'*. Two examiners (from the control group) felt that the demanding time scales of marking were negative: *'I always feel that the standardisation process is somewhat delayed following the taking of the paper and the deadlines set (for marking) are frequently unrealistic', 'a big problem in achieving standardisation is the shortage of time in which to complete the marking. This places great pressure on examiners to complete marking and possibly accuracy is sacrificed'*. On the other hand a third examiner (also from the control group) felt that: *'The more demanding the time scale the more likely that marking will be accurate'*.

As part of the control group two examiners said that they missed the discussion from a co-ordination meeting. Two others (from the non-hierarchical group) felt that discussion in the research project co-ordination meeting was an advantage an examiner in the non-hierarchical co-ordination group said: *'A good airing of a wide range of views. I certainly learnt a bit more about Khrusher! Always a pleasure to meet and exchange views'*. This is positive as it indicates that there was a noticeable difference between the levels of discussion in the non-hierarchical and hierarchical groups and that some examiners preferred this approach.

Discussion and Conclusions

In summary the research shows that all aspects of standardisation are considered by examiners to be important, particularly the mark scheme and the co-ordination meeting. It shows that the levels of attainment required to gain marks are communicated by discussing mark schemes in relation to exemplars, which fits with the views of Wolf (1995) and Hall and Harding (2002). The information gained from the questionnaire must be interpreted in the light of the

research literature and particularly in the light of the results of the two experiments reported here. The examiners thought that co-ordination meetings are useful, and some said essential. However the co-ordination meetings made no statistically significant difference to the reliability of marking in the experiment. But all the examiners had remarked another examination paper from the History GCSE in the last session, and therefore the result should not be generalised to all situations. It could be that co-ordination meetings are more important for new examiners and when significant changes are made to specifications. It seems that examiners value the opportunity in the co-ordination meeting to develop/reinforce a community of assessment practice and learn about the application of the mark scheme. It might be that the 'usefulness' of the co-ordination meeting is that it gives examiners a feeling of being part of a team, boosts confidence and gives the examiners some feedback. The results of this research should not be used to argue that co-ordination meetings are unnecessary.

The research literature suggests that exemplar scripts and discussion between examiners are important in facilitating marking. The examiners did not think these aspects of standardisation were as useful as the mark scheme and the co-ordination meeting, and they said the exemplar scripts and the discussion simply facilitated the understanding and application of the mark scheme. One of the points which is often made in the literature about assessment using banded mark schemes or criteria is that not everything can be written down and that some understanding of how to apply the mark scheme/criteria will remain tacit amongst a community of practitioners. But the examiners pointed out that what is written down in terms of the mark scheme and how it is written, i.e. clearly, is actually very important. This fits with the literature suggesting that the mark scheme has a standardising effect for members of an appropriate community of practice. Additionally examiners thought that exemplar scripts should be annotated and related to the levels in the mark scheme. Study 1 illustrates that the way that the exemplar scripts are tied to the mark scheme affects the severity of marking.

The principle of a community of practice would be that the application of a mark scheme can be negotiated between PEs and Assistant Examiners so that there is a shared understanding and ownership. Three examiners did mention that the mark scheme could be adjusted at the co-ordination meeting. Indeed in this study the mark scheme was adjusted at the co-ordination meeting when, in the non-hierarchical co-ordination meeting, it was changed according to the wishes of the examiners and the PE. However nine examiners mentioned that the co-ordination meeting makes the PE's interpretation of the mark scheme clear and eight said that discussion with someone in the hierarchy (PE or Team Leader) is useful. On the other hand the same number of examiners found discussion with other Assistant Examiners useful. Additionally Table 3 shows that examiners found discussion in a non-hierarchical co-ordination meeting more useful than discussion in a hierarchical co-ordination meeting (but not significantly so) and there was no significant difference between the hierarchical and non-hierarchical co-ordination meetings in terms of the accuracy of marking. The control group found discussion significantly more useful than the examiners at the hierarchical co-ordination meeting. In other words examiners would rather have a hierarchical discussion than no discussion and there is some preference for non-hierarchical rather than hierarchical discussion in a co-ordination meeting, although the type of discussion does not affect inter rater reliability.

Deferring to the Team Leader and/or PE fits with the general intention of a co-ordination meeting - to bring the Assistant Examiners' marking in line with that of the prime marker (PE). Although this is the aim of a co-ordination meeting there is literature which suggests that this cannot be achieved in the context of testing English as a foreign language, e.g. McNamara (1996). Some co-ordination meetings with a small number of examiners and/or some Team Leader meetings are conducted along the principles of a community of practice, where decision making is a non-hierarchical process of consultation and discussion. (Team Leaders meetings are where the mark scheme is tested by Team Leaders and adjusted where necessary before large co-ordination meetings). But it is difficult to include non-hierarchical decision making and consultation in large co-ordination meetings of say 150 examiners. In the communities of practice literature it is argued that more learning takes place when meaning is negotiated in a flat hierarchy. This was operationalised in study 2 through a non-hierarchical or more democratic style of decision making in a standardisation meeting. It appears from the responses to the questionnaire that examiners feel that meaning of the mark scheme can be satisfactorily negotiated in a hierarchical structure and that direct contact with the Principal Examiner is appreciated. In other words examiners want to air their views to senior examiners and ask them queries so

that the meaning of the mark scheme is clarified, and that non-hierarchical discussion and decision making is appreciated.

The comments made by the examiners suggest that at least some of them are engaged in a deep process of learning. That is they want to know 'how' and 'why' the exemplars relate to the level descriptors. Learning is part of a community of practice. But the notion of deep learning is from a different theory of learning, see Marton and Säljö (1976a and b), Biggs (1987, 1993), Entwistle, (1981) and Ramsden (1992). This suggests that other theories of learning should be considered as well as communities of practice to understand how examiners learn to apply a marking scheme.

Given these conclusions it is suggested that future research should focus on developing communities of practice amongst examiners and those involved in national assessments. For example, it might be advantageous to develop communities of assessment practice amongst teachers who are marking GCSE and A level coursework. The community should negotiate the meaning of the mark scheme with the senior examiners. A further project might investigate the advantages and disadvantages of co-ordination meetings for teachers who mark coursework (AQA has mandatory co-ordination meetings for teachers marking coursework, OCR does not). Although the co-ordination meetings in the research reported above did not affect the accuracy of marking it is obviously a valued part of the standardisation process for GCSEs. At the moment there are suggestions that national examinations might be taken by candidates on computers and that examiners might mark the answers on screen. This has many implications, one of which is that some of the communication between examiners might be conducted using secure websites or other means. Given these developments the advantages and disadvantages of electronic and face to face communities of assessment practice might be investigated.

References

Baird, J. Greatorex, J., and Bell, J. F. (2002) *What makes marking reliable? Experiments with UK examinations.* A paper presented at the International Association for Educational Assessment Conference, 1 - 6 September 2002, Hong Kong, China.

Barrett, S. (2000) *HECS LOTTO: Does Marker Variability make examinations a lottery?* Division of Business and Enterprise, University of South Australia. www.aare.edu.au/99pap/bar99789.htm

Biggs, J. (1987) *Student Approaches to Learning and Studying* Hawthorn, Vic: Australian Council for Educational Research

Biggs, J. (1993) What do inventories of students' learning process really measure? A theoretical review and clarification, *British Journal of Educational Psychology*, 83, 3-19.

Britton, J. N., Martin, N. C. and Rosen, H. (1966) *Multiple Marking of English Compositions An account of an experiment*, Schools Council Examinations Bulletin No 12, Her Majesty's Stationery Office.

Entwistle, N (1981) *Styles of Learning and Teaching; an integrated outline of educational psychology for students, teachers and lecturers* Chichester: John Wiley (0 471 10013 7).

Furneaux, C. and Rignall, M. (2000) *The effect of standardisation-training on rater-judgements for the IELTS Writing Module*, School of Linguistics and Applied Language Studies, University of Reading.

Hall, K. and Harding, A. (2002) Level Descriptions and Teacher Assessment in England: towards a community of assessment practice, *Educational Research*, 44 (1) 1-16.

Hartog, P. & Rhodes, E. C. (1935b) *The Marks of Examiners*, London. In Black, E L (1962) 'The marking of GCE scripts', *British Journal of Educational Studies*, 11, 61-71.

Konrad, J. (1998) *Assessment and Verification of National Vocational Qualifications: policy and practice*. Education-line www.leeds.ac.uk/educol/index.html.

Lenney, E., Mitchel, L., & Browning, C., (1983) The effect of clear evaluation criteria on sex bias in judgements of performance. *Psychology of Women Quarterly*, 7 (4), Summer, 313-327.

Lunz, M. E. and O'Neill, T. R. (1997) *A longitudinal study of judge leniency and consistency*. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, Illinois.

Lunz, M.E., Wright B.D. and Linacre, J.M. (1990) Measuring the impact of judge severity on examination scores. *Applied Measurement in Education* 3, 331-45.

Marton, F. and Säljö (1976a) On Qualitative Differences in Learning — 1: Outcome and Process, *British Journal of Educational Psychology*, 46, 4-11.

Marton, F. and Säljö (1976b) On Qualitative Differences in Learning — 2: Outcome as a function of the learner's conception of the task, *British Journal of Educational Psychology,* 46, 115-27

McNamara, T. (1996) *Measuring second language performance,* Harlow: Longman.

Qualifications and Curriculum Authority (2002) *GCSE, GCSE in vocational subjects, GCE, VCE and GNVQ code of practice 2002/3*, Qualifications and Curriculum Authority, London. www.qca.org.uk

Ramsden, P. (1992) *Learning to Teach in Higher Education,* London: Routledge (0-415-06415-5).

Shaw, S. (2002) The effect of standardisation training on rater judgement and inter-rater reliability for the revised CPE writing paper 2, *Research Notes 8*, May 2002

Stahl, J. A. and Lunz, M.E. (1991) *Judge performance reports: media and message*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Weigle, S. (1998) Using FACETS to model rater training effects, *Language Testing*, 15 (2) 263-287.

Wenger, E. (1998) *Communities of Practice Learning, meaning and identity*, Cambridge University Press: Cambridge.

Wigglesworth, G. (1993) Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing* 10(3): 305-35.

Wiliam, D. (1998) *Enculturating learners into communities of practice: raising achievement through classroom assessment*, Paper presented at the European Conference for Educational Research, University of Ljubljana, Slovenia, September 17th to 20th.

Wolf, A. (1995) *Competence Based Assessment*. Open University Press: Buckingham.

APPENDIX A QUESTIONNAIRE: RESEARCH STUDY ON MARKING

To Jackie Greatorex, RED, UCLES, 1 Hills Rd, Cambridge, CB1 1DR.

From

1. How useful do you think each of the following are for standardising marking?

i) Mark scheme

Not useful	1	2	3	4	5	6	7	Useful

Please circle one number and explain your answer below

_____

_____

_____

_____

_____

_____

_____

ii) Co-ordination meeting

Not useful	1	2	3	4	5	6	7	Useful

Please circle one number and explain your answer below

_____

_____

_____

_____

_____

_____

_____

PTO

iii) Discussion with other examiners

Not useful	1	2	3	4	5	6	7	Useful

Please circle one number and explain your answer below

_____

_____

_____

_____

_____

_____

iv) Exemplar scripts and associated marks awarded by the Principal Examiner

Not useful	1	2	3	4	5	6	7	Useful

Please circle one number and explain your answer below

_____

_____

_____

_____

_____

_____

_____

2. Do you think the marking standardisation process was better for the operational examination? (delete as appropriate) Yes/No If so, why?

_____

_____

_____

_____

_____

_____

3. Do you have any further comments on your experience of participating in this study?

_____

_____

_____

_____

THANK YOU FOR COMPLETING THE QUESTIONNAIRE.

APPENDIX B: RESULTS (QUESTIONS 1I, 1II, 1III AND 1IV)

Frequency Tables

Table 5 Question 1i - Mark scheme

Response	Frequency	Percent	Cumulative Percent
3	2	5.7	5.7
4	1	2.9	8.6
5	3	8.6	17.1
6	9	25.7	42.9
7	20	57.1	100.0
Total	35	100.0	

Table 6 Question 1ii - Co-ordination meeting

Response	Frequency	Percent	Cumulative Percent
3	1	2.9	2.9
4	2	5.7	8.6
5	4	11.4	20.0
6	8	22.9	42.9
7	20	57.1	100.0
Total	35	100.0	

Table 7 Question 1iii - Discussion with other examiners

Response	Frequency	Percent	Cumulative Percent
1	1	2.9	2.9
2	4	11.4	14.3
3	1	2.9	17.1
4	5	14.3	31.4
5	6	17.1	48.6
6	7	20.0	68.6
7	11	31.4	100.0
Total	35	100.0	

Table 8 Question 1iv - Exemplar scripts

Response	Frequency	Percent	Cumulative Percent
3	3	8.6	8.6
4	6	17.1	25.7
5	8	22.9	48.6
6	10	28.6	77.1
7	8	22.9	100.0
Total	35	100.0	