The Effect of Marker Background and Training on the Quality of Marking in GCSE English

Introduction

In the UK, the selection of markers for national examination systems is largely a matter of custom and practice. The criteria used by the Assessment and Qualifications Alliance (AQA) are comparable to those used by other UK awarding bodies. These are that examiners should have suitable academic qualifications (usually a relevant degree or equivalent) and at least three terms' teaching experience which should be recent and relevant. These selection criteria have face-validity, as it would seem appropriate to insist upon a relevant educational background and teaching experience at the appropriate level for the marking of examinations. Indeed the code of practice governing UK awarding body procedures (QCA, 2007) demands that examiners must have relevant experience in the subject but does not explicitly discuss the nature of this experience.

The proliferation of examining and the introduction of computer-based assessment have meant that the search for an empirically supported definition of 'relevant experience' has taken on new importance. Examiners are in short supply and e-marking technology has provided the facility for individual items within an examination to be marked separately, by individuals with different backgrounds. Investigations of the relationship between individual differences and marker reliability are crucial in determining examiner recruitment practices. A number of studies have attempted to identify factors that might allow the identification of those examiners who are likely to mark most reliably and those who are likely to require additional training or monitoring. These studies are reviewed below.

The relationship between examiner background and marking performance

Research suggests that compared to experienced markers; inexperienced markers tend to mark more severely and employ different rating strategies (Ruth and Murphy, 1988; Huot, 1998; Cumming, 1990; Shohamy, Gordon and Kraemer, 1992; Weigle, 1994, 1999). Ruth and Murphy (1988) reported a study that revealed a tendency for trainee teachers to mark essays more severely than experienced markers, though the differences were not significant. They suggested that the markers' background determined distinctly different frames of reference for judging the essays. Similarly, Weigle (1999) reported that inexperienced examiners were more severe than experienced examiners. She found that prior to training, inexperienced markers could be significantly more severe than experienced markers depending on the essay title, but after training the differences in severity disappeared. She suggested that her results *"underscore the complexity of the relationship between rater background, the scoring rubric, the prompt, and rater training in writing assessment."* (p.171)

Not all studies have replicated the relationship between inexperience and marking severity. Myford and Mislevy (1994) studied the Advanced Placement examination in Studio Art in the US. They attempted to identify background variables, including years of teaching experience, which might predict marker severity but found that the variables studied had a negligible impact on predictions of marker severity. Further, Meyer (2000a, 2000b), investigating marking in AQA's GCSE English Literature and Geography, found that length of examiner experience and a senior examiner's rating of the examiner's performance (from A - consistently excellent, to E - unsatisfactory not to be re-employed) rarely proved useful as predictors of whether an examiner's marks would require adjustment to correct for severity or generosity.

While there is some evidence of an association between marker experience and severity, studies have failed to differentiate the effects of teaching and examining experience. Moreover, in large scale testing programmes concern is often focused on inconsistency rather than severity in marking. Variations across examiners in marking severity can be accounted for by adjusting candidates' marks and this is common practice in UK awarding bodies (Baird and Mac, 1999). However, mark adjustment can only be used where the examiner has been consistently severe or lenient. It is of no help when markers are inconsistent in their application of the mark scheme. So marking inconsistency is a much greater threat to the reliability of the marks awarded to candidates. Evidence of an association between marker background and marking consistency will now be reviewed. It is, however, ambiguous, and studies investigating this relationship have generally failed to tease out the effects of markers' subject knowledge, teaching and marking experience on marking consistency.

Ecclestone (2001) carried out a case study of nine university lecturers who double-marked 45 dissertations between them over two years. Discrepancies between grades were moderated at a one-day moderation meeting, and the external examiner saw a sample of dissertations. Rough distinctions between the lecturers were made according to length of experience in assessing the programme and of other degree and Masters' level work. The lecturers were classified as novice, competent or expert markers. Following moderation, the novices had fewer changes to their marks than the competents and experts, with the competents having more than the other two groups. However, experts had more changes that resulted in the degree grade being altered by a whole degree class while competents had more changes to their marks but within the same degree classification.

Also working in the higher educational context but in the US, Michael, Cooper, Shaffer and Wallis (1980) compared marks of two English essays given by university professors of English (defined as expert markers) and professors of other disciplines (defined as lay markers). The reliability indices were slightly higher for marks provided by either individual experts or pairs of experts than for those provided by lay readers or pairs of lay readers, but the differences were small enough for the authors to conclude that the reliability of the two groups was nearly comparable. Differences in reliability were greater between essay questions than between the types of marker suggesting that reliability was more a function of the type of question or of variations in the average ability level of the examinee samples than of the expertise of the markers. This pattern of findings was repeated for measures of concurrent validity[1] of the essay evaluations. Expert markers' evaluations had slightly higher validity than those of lay markers, but the variation in validity associated with the different essay questions were far greater.

Shohamy, Gordon, and Kramer (1992) studied marker reliability in the assessment of English as a foreign language (EFL) among markers who were either professional, experienced EFL teachers or lay people (native English speakers). Half were trained in one of the three marking procedures used (holistic, analytic and primary trait scoring). Relatively high inter-rater reliability was achieved by the four groups of markers (trained/professionals, untrained/professionals, trained/lay and untrained/lay), irrespective of the type of training received, but the overall reliability coefficients were higher for trained markers than they were for the untrained ones.

[1] As assessed by three criterion measures: Diagnostic Test of Written English; Test of Standard Written English; and grade point average across all college or university courses.

Therefore, training appeared to have significant effect on marking, but no such effect was found for markers' background. The findings suggested that markers are able to mark reliably, regardless of background as long as they are given intensive procedural training. As Shohamy *et al* note,

"*the practical implication of this finding is that decision makers, in selecting raters, should be less concerned about their background, since that variable seems not to increase reliability. More emphasis, however, should be put into intensive training sessions to prepare raters for their task*." (p. 31)

In another study of English assessment but in Australia, Lumley, Lynch and McNamara (1994) had doctors and trained Occupational English test raters rate the overall communicative effectiveness of 20 candidates taking the Occupational English test. There was no difference between the two groups of raters in terms of severity, although if anything the doctors were slightly more lenient. Moreover, all but one of the doctors interpreted the scale consistently with the experienced raters.

Brown (1995) investigated rater background factors in assessment on the Japanese Language test for Tour Guides, an oral test measuring Japanese Language skills of Australian tour guides. Assessors were either from the tourist industry (this was preferred) or they were experienced teachers of Japanese as a foreign language. Overall the occupational background had no effect on rating severity or perhaps more interestingly consistency. There was, however, greater variability in levels of severity among the non-teacher group. There were also differences between the groups at the level of particular criteria: teachers were harsher in ratings of grammar, expression, vocabulary and fluency, whereas industry raters gave harsher ratings of pronunciation. There was also some variation in severity across task type and in the way raters interpreted the ratings scales, for example teachers were less prepared to award very high or low scores. Nonetheless, the differences were not such as to suggest that the two groups differed in their suitability as raters.

Pinot de Moira (2003) studied the relationship between examiner background and marking reliability across seven AQA GCE subjects. She defined reliability as the difference between senior examiner and assistant examiner mark; the absolute difference between senior examiner and assistant examiner mark; whether an adjustment had been made to the assistant examiner's marks and a rating of the examiner's performance (from A - consistently excellent, to E - unsatisfactory not to be re-employed). She found that the composition of an examiner's script allocation in terms of centre type had far more influence on accuracy than accessible aspects of an examiner's background, such as years since appointment. The only personal characteristic found to be significant in explaining examiner reliability was the number of years of marking experience. Royal-Dawson (2004) pointed out however that this characteristic was confounded because reliable examiners are engaged year after year and poor markers are not, so quality of marking and length of service are not mutually exclusive.

Some studies have focused specifically on whether teaching experience is a necessary requirement for accurate marking. Working in the US, Powers and Kubota (1998a) investigated whether individuals not involved in post-secondary teaching could accurately mark essays written by college students seeking admission to graduate programmes in business management. To this end, they compared the quality of marking of experienced and inexperienced examiners.

The experienced markers had previously participated in the holistic scoring of essays for one or more Educational Testing Service (ETS) administered testing programs. All had graduate degrees and taught in university-level courses involving critical thinking skills or writing. The inexperienced group either did not have graduate degrees or were not currently teaching college level courses involving critical thinking skills or writing and had no experience of the holistic scoring of essays. All had a baccalaureate degree.

Essays were marked before and after training. After training, inexperienced markers especially, improved significantly in their ability to assign 'correct' scores. However, several of the inexperienced markers were as accurate as the experienced markers even before the training. Powers and Kubota concluded that there were 'few significant relations between background and accuracy' and that the current pre-requisites for ETS essay markers would automatically disqualify a proportion of potential markers, who could, after training, mark accurately.

Powers and Kubota (1998b) extended this study to a second kind of essay writing prompt – 'analysis of argument' which is used to select candidates for graduate programs in management. As in the previous study, the results suggested that inexperienced markers without the currently required credentials could be trained to score 'argument' essays with a high degree of accuracy. They also collected logical reasoning scores for the markers. The results suggested a possible link between logical reasoning and marking accuracy. It is unfortunate that Powers and Kubota's design did not extricate teaching experience and subject knowledge as it is likely that these are differentially important in marking performance.

In the UK Royal-Dawson and Baird (Royal-Dawson, 2004; Royal-Dawson and Baird, in preparation) explored whether it is necessary for a marker of Key Stage 3 English to be a qualified teacher with three years' teaching experience. They examined the marking reliability of four types of markers with an academic background in English but different amounts of teaching experience: English graduates, PGCE graduates, teachers with three of more years' teaching experience and experienced examiners. Reliability was defined in a number of ways: the correlation between the marks awarded to the 98 scripts by the Lead Chief Marker and the marker; the agreement between the levels assigned to a pupil by a marker compared to those assigned by the Lead Chief Marker; the frequency of administrative errors. Overall, there was little difference in the marking reliability of the different types of marker. There were more or less accurate markers in each of the groups, but no group had more or fewer accurate markers than any other. Marking reliability, as defined by the correlation between each marker and the Lead Chief Marker, indicated that some teaching experience was a contributing factor to higher reliability estimates on some tasks but not on others. There was no difference in lenience or severity between the marker groups except on a sub-test for reading where the experienced markers were more lenient than the other marker groups. They concluded that the criterion of teaching experience could be relaxed to allow markers with graduate-level subject knowledge to mark Key Stage 3 English tests.

To summarise, research conducted across countries, test types, mark schemes, subject areas and skills; using a variety of methodologies; analysing data from designed studies and operational data; has failed to find a consistent association between aspects of markers' background and marking reliability. One of the main criteria used by awarding bodies for evaluating the employability of an examiner is relevant classroom experience. However, there is little empirical evidence for a relationship between examiner teaching background and marking reliability. If teaching experience is not the key criterion for judging the suitability of potential expert examiners, on what basis should applicants be judged? Is subject knowledge rather than teaching experience key to being reliable marker? Or, with the right training, can anyone mark reliably?

Current Study

This study explores the reliability with which individuals with distinctly different education, teaching and examining backgrounds mark GCSE English. This will provide an opportunity to attempt to replicate the finding that classroom experience is not a pre-requisite of reliable marking in Key Stage 3 English (Royal-Dawson, 2004; Royal-Dawson and Baird, in preparation). The design goes further though in attempting to disentangle the effects examining experience, teaching experience and subject knowledge. It is likely that the relationship between markers' background and marking reliability will vary with the kind of item being marked (as was clearly demonstrated in the study of Key Stage 3 English marking). For example, an individual with no subject knowledge may be able accurately to mark short answer questions but not essay questions. To enable investigation of this possibility, participants were required to mark a mixture of items requiring both short and longer responses. Hence, the findings may support the selection and employment of individuals with non-teaching backgrounds as examiners in subjects where there is an examiner shortage and will inform the development of guidelines as to the suitability of different items types for e-marking by different types of marker, expert or general (clerical), for example.

Methodology

Four groups of participants were recruited to mark the same two hundred 2005 GCSE English A, Higher tier, Paper 1, Section A part-scripts. Part, rather than whole, scripts were marked to increase the variety of work marked by participants. They marked one section of the question paper, which included two questions: the first required two relatively short answers and one slightly longer answer; the second required two longer answers (see Figure 1 for a summary of the question paper section). GCSE English was considered a suitable subject because historically there is evidence of relative unreliability in marking, adjustments are applied to the marking, for example), the question papers include a variety of items possibly requiring different levels of skill and the subject is not so specialist as to make reliable marking by non-English graduates impossible.

Figure 1. A summary of the section of the question paper

Candidates were asked to refer to
1: An extract from Bill Bryson's book *Why No One Walks*
2: A car advertisement taken from the *Guardian* called *Gadgets for the Girls*

1a) What surprises Bryson about the way Americans Live? **(3 marks)**
1b) What method does Bryson use to entertain the reader? **(4 marks)**
1c) Compare the views in Item 1 with the views about cars in Item 2. **(6 marks)**

2a) How does the use of language in the advertisement make the car seem desirable? **(8 marks)**
2b) How effective are the pictures in helping support the claims made for the car in the written text? **(6 marks)**

The part-scripts had been marked by the Principal Examiner for the question paper. The Principal Examiner was responsible for setting the paper and mark scheme and standardising its marking during its use as a live examination paper.

The groups of participants are described in Table 1. They were selected to enable the relative importance of previous examining experience, subject knowledge and teaching experience to marking reliability to be assessed. A short screening questionnaire ensured that participants had the requisite amount of teaching experience and subject knowledge to qualify for inclusion. For example, the English undergraduates and the undergraduates from other disciplines had negligible or no teaching experience. The experienced markers had previously marked a different GCSE English question paper but had no previous experience of marking the paper used in this study.

Table 1 Groups of markers participating in the study

	Marking experience	Subject knowledge	Teaching experience	N
Experienced GCSE English A Paper 2 markers	high	high	high	97
PGCE English undergraduates	low	high	some	81
English/Linguistic undergraduates	none	high	none	99
Other discipline undergraduates	none	low	none	82

The procedure is summarised in Figure 2. The study was conducted in a marking centre. Initially participants marked a first batch of 100 part-scripts by applying the mark scheme (no marking standardisation training had been received). They then received the current training and standardisation procedures for GCSE English A Paper 1 markers. Seven exemplar scripts were used in the training. After participants had marked each of the seven scripts the Principal Examiner discussed the 'standardised' marks with the group. Participants then marked another batch of 99 part-scripts. Scripts were randomly sampled from over 220,000 scripts marked during the summer 2005 examination period. Since research suggests that marking reliability varies with the quality of work (Pinot de Moira, 2003), care was taken to ensure that the samples covered the full mark distribution. Scripts were cleaned using a scanner and filter to remove the original examiners' marks.

Figure 2. A summary of the procedure

Day 1:
Marked 100 GCSE English A paper 1H part scripts

Day 2:
Standardisation training conducted by the Principal Examiner

Day 3:
Marked another batch of 100 GCSE English A paper 1H part scripts

Results

Operationalisation of quality of marking

There is a variety of methods of assessing quality of marking. These include:

- The relative severity/leniency of the marking;
- The absolute difference between the mark given by the marker and the estimated 'true' mark;
- The correlation between the marks given by the marker and the estimated 'true' mark.

There are other important aspects to the quality of marking not investigated in this study; the administration of the marker, for example. A marker who is reliable but who does not return their work on time wouldn't be highly regarded.

There is also more than one conceptualisation of 'true' mark. These include:

- The mark given by the Principal Examiner, who is the most senior examiner of the question paper. 'True' mark is operationalised by UK awarding bodies in this way;
- A consensual view of the 'true' mark, for example the mean mark allocated by all markers. This view of 'true' mark is similar to that embodied by classical test theory, that is, the mark given by the pooled judgement of an infinite number of markers (Spearman, 1904a, 1904b, 1927).

Taking both a consensual and a hierarchical approach to estimating the 'true' mark will allow findings to be generalised to assessment systems that employ either approach. It also guards against the possibility of the study's conclusions being influenced by error in the Principal Examiner's marking of the scripts.

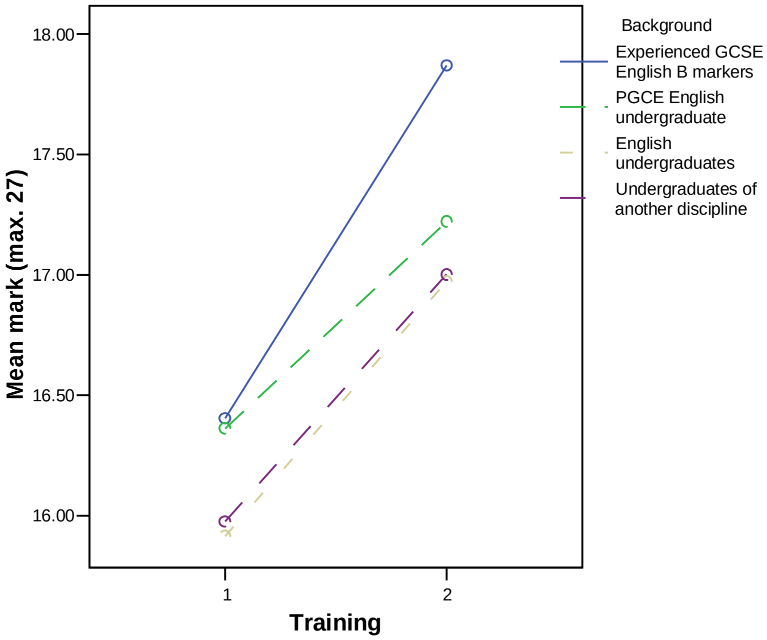
These approaches were used to investigate the quality of marking of the groups of participants before and after marker standardisation training (using two-way mixed ANOVAs). Unfortunately it is inappropriate to use correlations with restricted mark ranges e.g. Item 1a has a mark range of 0-3. This may reduce the statistical power of these analyses leading to effects being missed (Type II errors) or may lead to the detection of spurious effects (Type I errors). With this in mind, the correlation between the participants' marks and the estimated 'true' mark has only been investigated at the level of part-script (out of 27 marks) and for item 2a (out of 8 marks). In the interests of parsimony, investigations into the relative severity/leniency of marking are also reported at part-script level only.

Part-script total (27 marks)

Severity of marking

Before training the marking of Examiners and PGCE students was approximately half a mark more generous than that of the undergraduates but this difference was not significant. Following training all groups were more generous. This effect was not significantly different across the groups (see Figure 3).

Figure 3 The effect of marker background and training on the marks awarded by the participants to part-script total



Training also had the unexpected affect of reducing the spread of marks awarded. This can be seen from the standard deviation of marks before and after training (see Table 2). The effect was greatest for the PGCE students and least for the Examiners' marking. This is unfortunate since an explicit function of training is to stretch the range of marks awarded so as to avoid compression of the final mark distribution and hence the grade boundaries. Indeed training materials distributed to senior examiners refer to the desirability of encouraging a spread of marks. It's likely that training does not have a substantial effect of reducing the spread of marks in live marking since there is no evidence of grade boundary compression in this paper (the judgemental boundaries being: A 41, C 31, D 23, the maximum mark being 54).

Table 2 The standard deviation of marks awarded before and after training

Background	Before training	After training	Change
Examiners	147.59	134.10	-13.49
PGCE students	180.12	129.40	-50.72
English undergraduates	193.01	154.13	-38.88
Undergraduates	166.75	148.90	-17.85

Reliability of marking

There was no significant effect of marker background on the absolute difference in marks awarded by the participants and either the consensual or the hierarchical estimation of the 'true' mark (see Figures 4 and 5). There was a significant positive impact of training on this measure of marking reliability. Training reduced the absolute difference between participants' marks and either estimation of the 'true' mark. This was the case no matter what the background of the participants.

Figure 4 The effect of marker background and training on the absolute difference in marks awarded by the Principal Examiner and the participants to part-script total

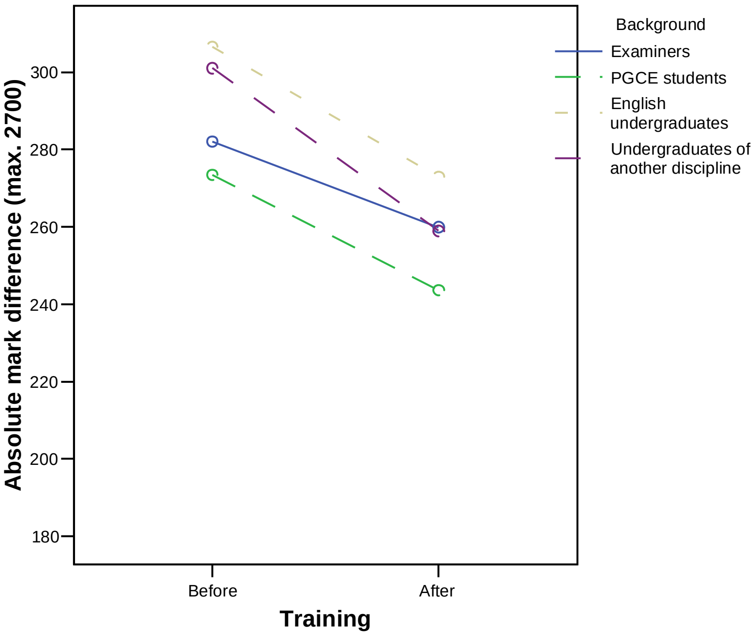
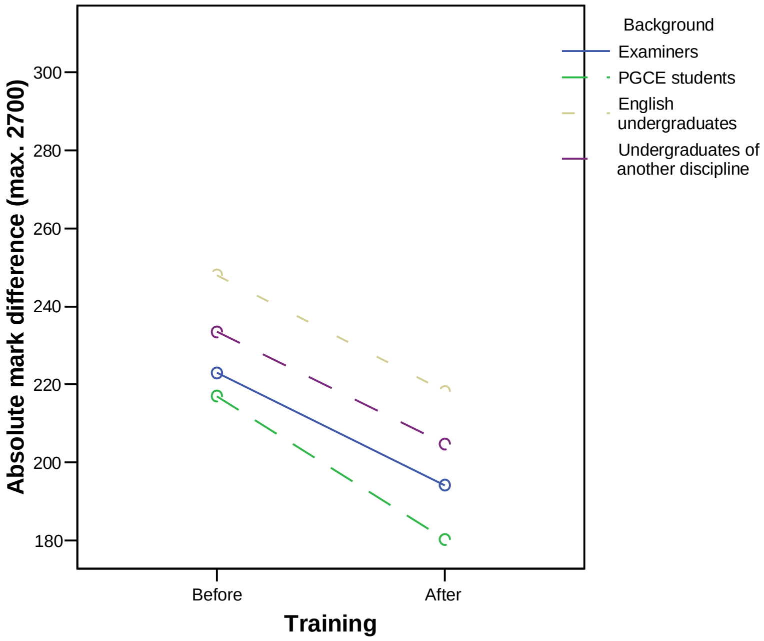


Figure 5 The effect of marker background and training on the absolute difference in the mean mark awarded to total part-script by all the participants and that awarded by individual participants



Since the part-script was marked out of 27 it was possible to examine the correlation between the participants' marks and the estimated 'true' marks. A Fisher transformation was applied to the correlation data to allow their use as dependent variables in ANOVA (Clark-Carter, 2006).

The effects of background were the same whichever estimation of 'true' mark was used. There was a significant effect of marker background on the correlation between the participants' marks and the 'true' marks. Tukey *post-hoc* contrasts showed that examiners marked this item significantly more consistently than both groups of undergraduates did. Further, the PGCE students marked more consistently than the undergraduates did. Note there was no significant difference in this measure of the reliability of marking of PGCE students and examiners.

The effects of training did, however, vary with the estimation of 'true' mark. Using a hierarchical definition, training had no significant impact on the marking of the undergraduates or English undergraduates, while it reduced the marking consistency of the examiners and PGCE students who had relatively high correlation coefficients prior to training (see Figure 6). On the other hand, using the consensual definition, training significantly improved the marking of the English undergraduates and undergraduates, who tended to mark relatively unreliably before training. Training had little impact on the marking consistency of the examiners but clearly reduced the quality of marking done by PGCE students (see Figure 7).

Hence, training both reduced the absolute mark difference from the estimated 'true' mark and for some groups of participants, reduced the correlation of their marking with the estimated 'true' mark. One might expect a reduction in absolute mark difference to be associated with an increase in the correlation. However, the effect of training on the spread of marks awarded explains this seemingly contradictory finding. Participants were more likely to award extreme marks before training. Training made them more cautious markers. This reduction in the spread of marks awarded impacted on the correlation with the estimated 'true' mark.

Figure 6 The effect of marker background and training on the correlation between the Principal Examiner's and participants' marking of the part-scripts

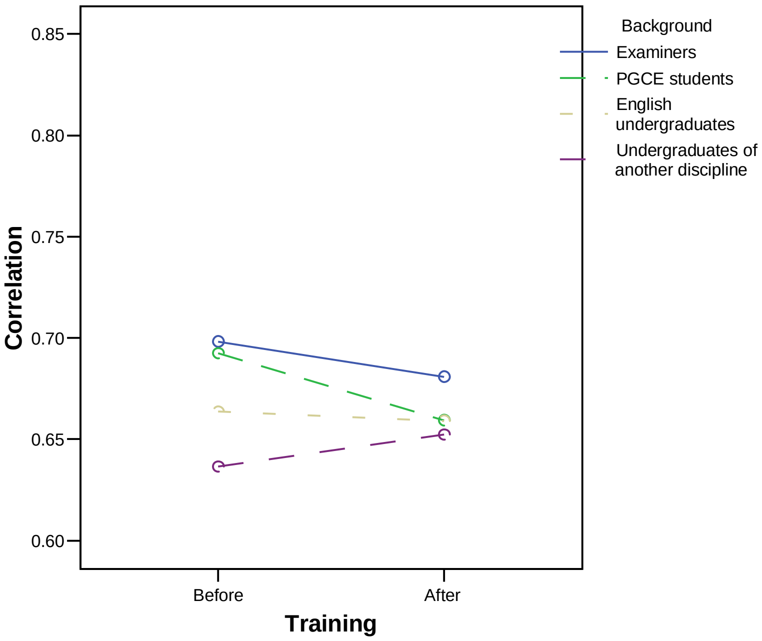
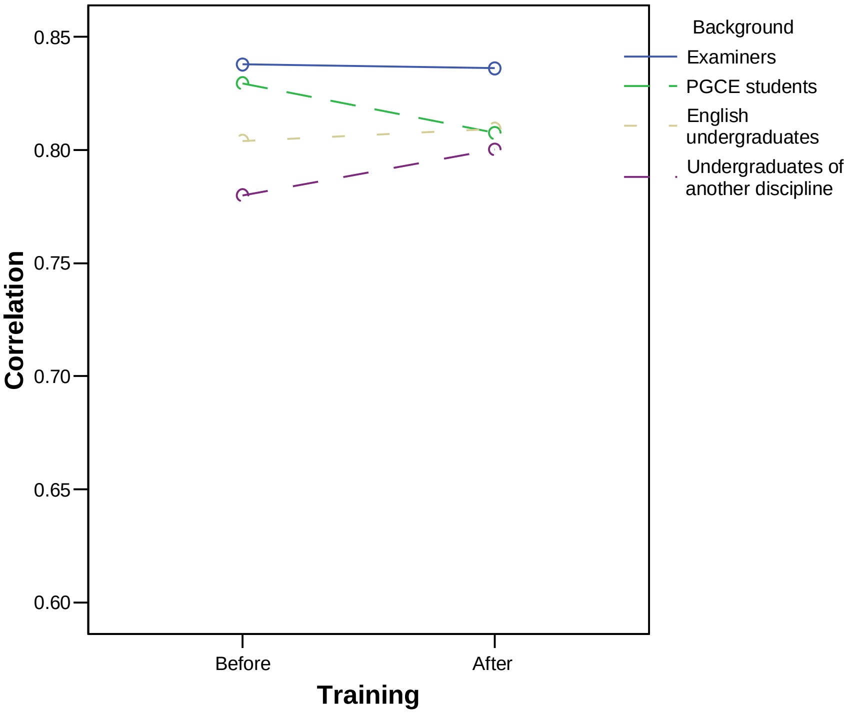


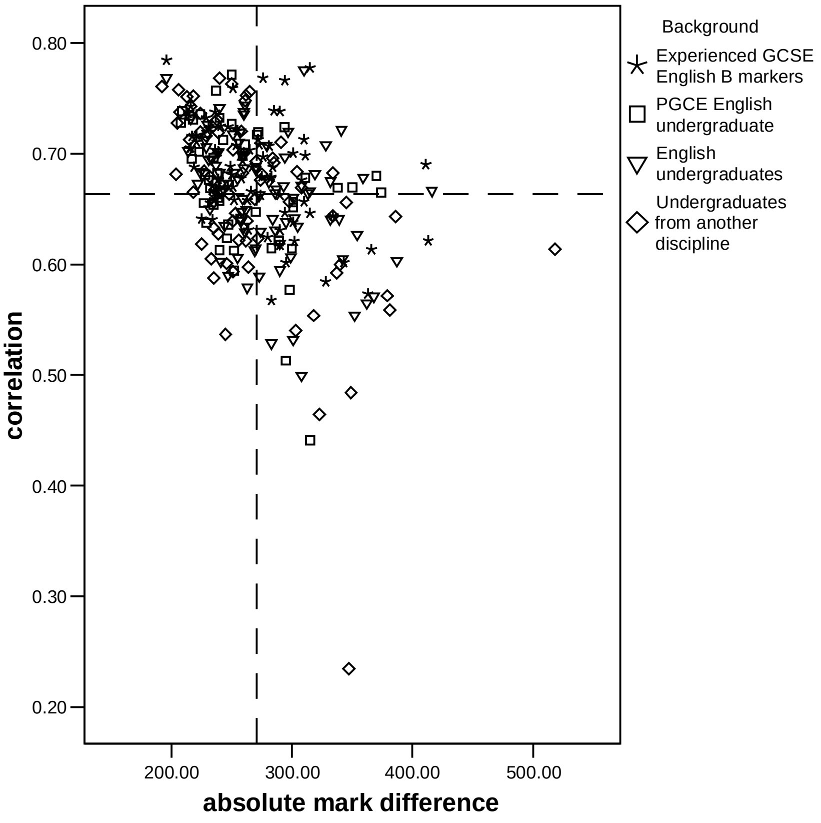
Figure 7 The effect of marker background and training on the correlation between the mean of marks awarded by all participants and the marks awarded by individual participants' to candidates' responses to total part-script



Reliable enough marking?

At a part-script level, the evidence suggests no significant difference in the quality of marking of PGCE students and examiners but the marking of undergraduates, English or otherwise, was significantly poorer than that of examiners. There were, however, some undergraduates who marked as well as the best examiners. This raises the question: what constitutes reliable enough marking? Using the hierarchical estimation of 'true' score, Figure 8 shows the both the correlation and absolute mark difference of the participants labelled by group. The lines represent the mean correlation and absolute maker difference of the *examiners*. One way of defining a 'good' marker, is an individual with a lower than average absolute mark difference and a higher than average correlation, that is those participants in the top left-hand quarter of the graph. The percentage of participants from each group defined as 'good' is remarkably similar; 43% of the Examiners, 43% of the PGCE students, 43% of the undergraduates from another discipline and 37% of English undergraduates. So while treating reliability as a continuum suggests an effect of marker background, using a categorical definition might lead to a different conclusion.

Figure 8 A scatter-plot of mean absolute difference in marks awarded by the Principal Examiner and the participants against the mean correlation between the Principal Examiner's and participants' marking of the part-scripts



Reliability of marking of items

Item 1a (3 marks)

There was no significant main effect of marker background on the absolute difference in marks between both the hierarchical and consensual estimation of 'true' mark and the marks awarded by the participants. Overall the groups of markers performed similarly well at marking this item.

The effect of training on reliability was not statistically significant when the hierarchal definition of 'true' mark was used. There was, however, a marginally significant interaction effect such that training had a detrimental impact on the marking reliability of the examiners and PGCE students (who marked relatively reliably before training) but a positive impact on the marking of the undergraduate groups (who marked relatively unreliably before training) (see Figure 9).

However, when the consensual definition was used training had a statistically significant positive effect, overall reducing the absolute mark difference from the consensual measure of 'true' mark. This effect was not statistically significantly different for participants with different backgrounds although Figure 10 shows that for the Examiners and to a greater extent the PGCE students marking deteriorated whereas for the undergraduates marking improved following training.

On balance there seems to be evidence to suggest that training did not have an expected positive impact on marking reliability for the examiners and PGCE students. While these groups were marking slightly more reliably than the undergraduate groups before training, the training brought their marking reliability to a level similar to that of the undergraduates.

Figure 9 The effect of marker background and training on the absolute difference in marks awarded by the Principal Examiner and the participants to item 1a

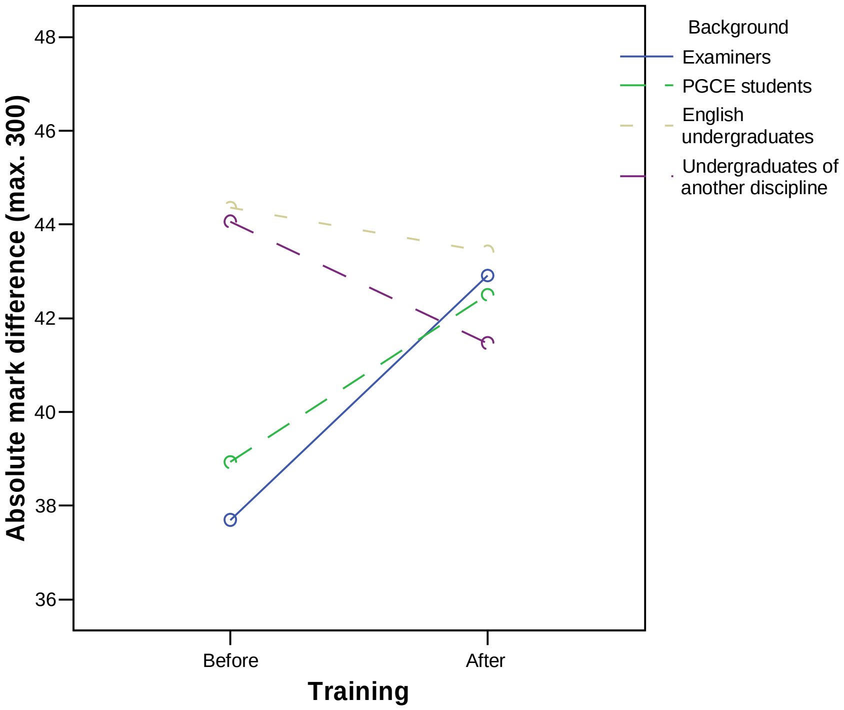
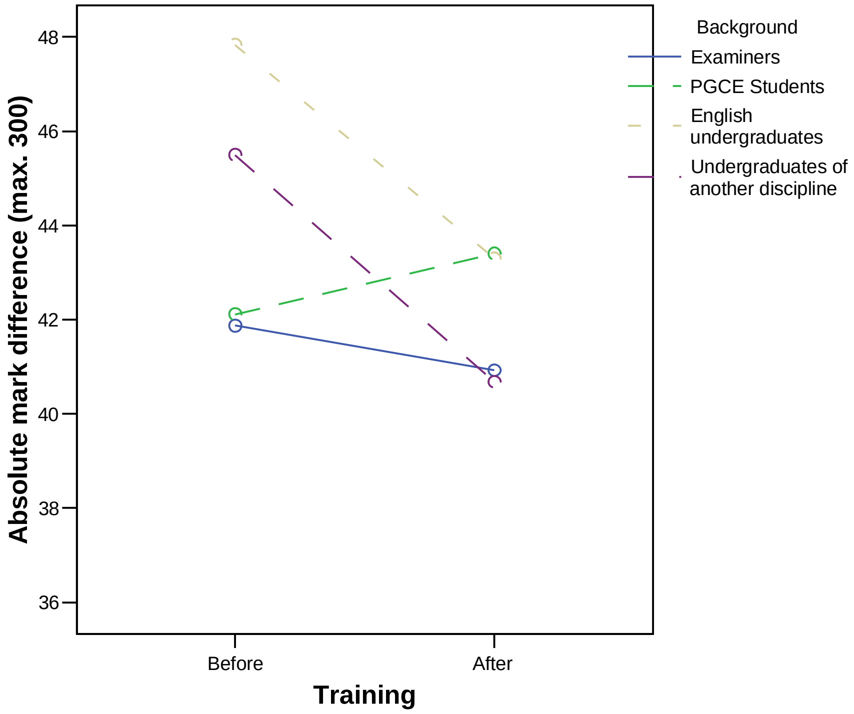


Figure 10 The effect of marker background and training on the absolute difference in the mean mark awarded to item 1a by all the participants and that awarded by individual participants



Item 1b (4 marks)

There was a significant effect of marker background on the difference in marks awarded to responses to this item by the Principal Examiner and the participants. PGCE students were significantly more reliable than the English undergraduates were. However this effect disappeared when the consensual definition of 'true' mark was adopted. Using this definition, there was no difference in reliability between the groups. Whichever 'true mark' was used, there was a significant positive impact of training on reliability which was equal across the groups of participants (see Figures X and X).

Figure 9 The effect of marker background and training on the absolute difference in marks awarded by the Principal Examiner and the participants to item 1b

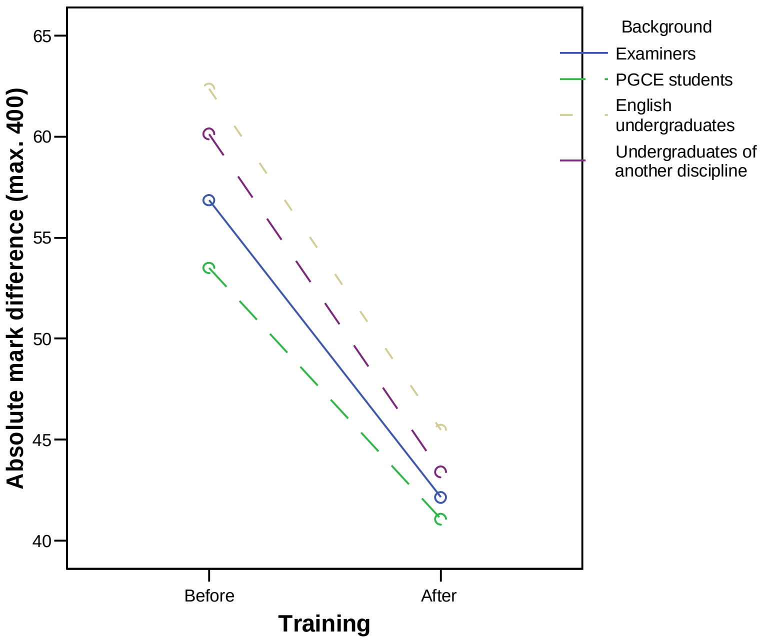
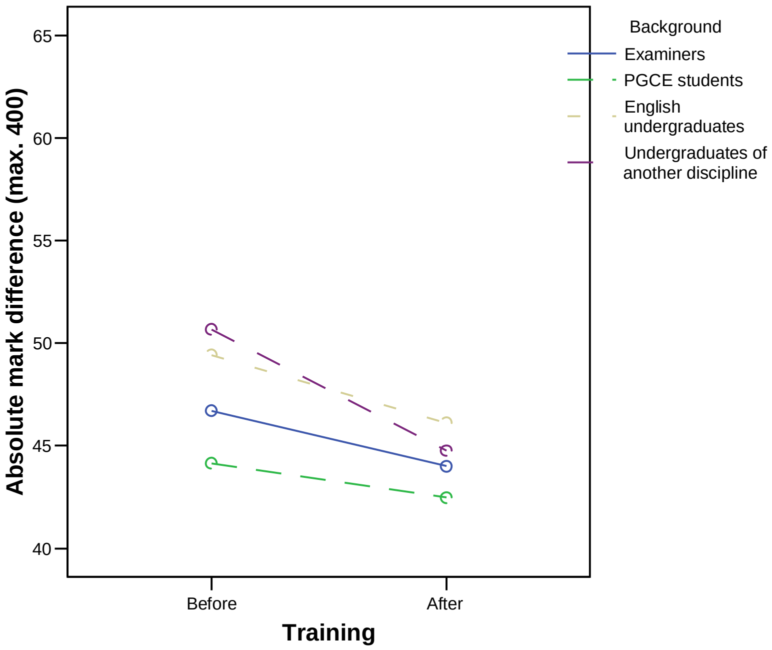


Figure 10 The effect of marker background and training on the absolute difference in the mean mark awarded to item 1b by all the participants and that awarded by individual participants



Item 1c (6 marks)

The findings did not vary according to which definition of ‘true’ mark was used. There was no significant effect of marker background on reliability and training had a significant positive impact which was not significantly different for participants with different backgrounds (see Figures 11 and 12).

Figure 11 The effect of marker background and training on the absolute difference in marks awarded by the Principal Examiner and the participants to item 1c

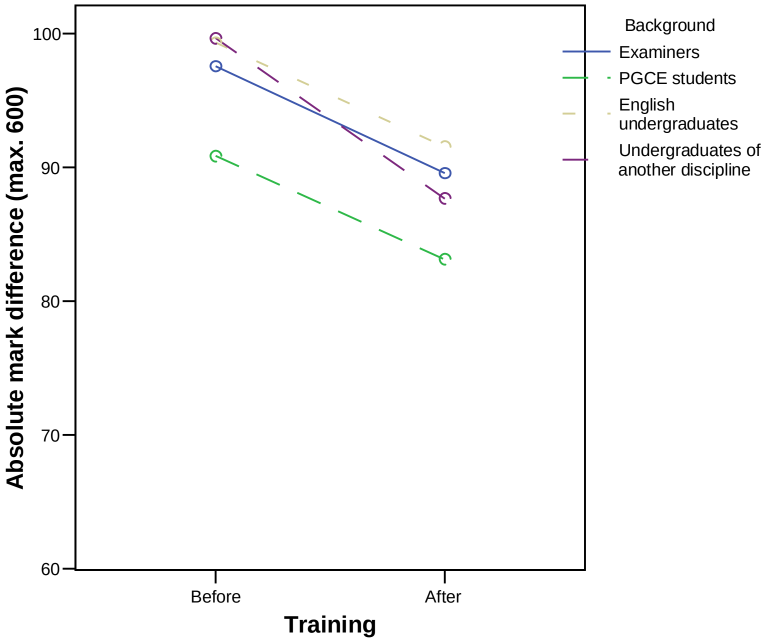
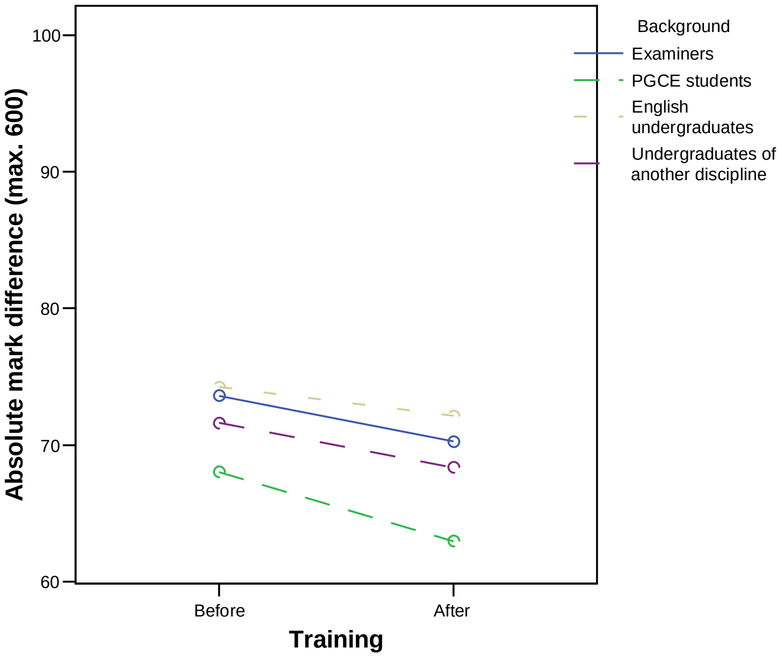


Figure 12 The effect of marker background and training on the absolute difference in the mean mark awarded to item 1c by all the participants and that awarded by individual participants



Item 2a (8 marks)

The findings were the same for both definitions of 'true' mark. There was a significant main effect of marker background on the absolute difference in marks. Tukey *post hoc* contrasts were non-significant but Figures 13 and 14 show that the examiners and PGCE students marked more reliably than the English undergraduates and undergraduates. There was a significant positive impact of training on reliability which did not interact with the background of the participants.

Figure 13 The effect of marker background and training on the absolute difference in marks awarded by the Principal Examiner and the participants to item 2a

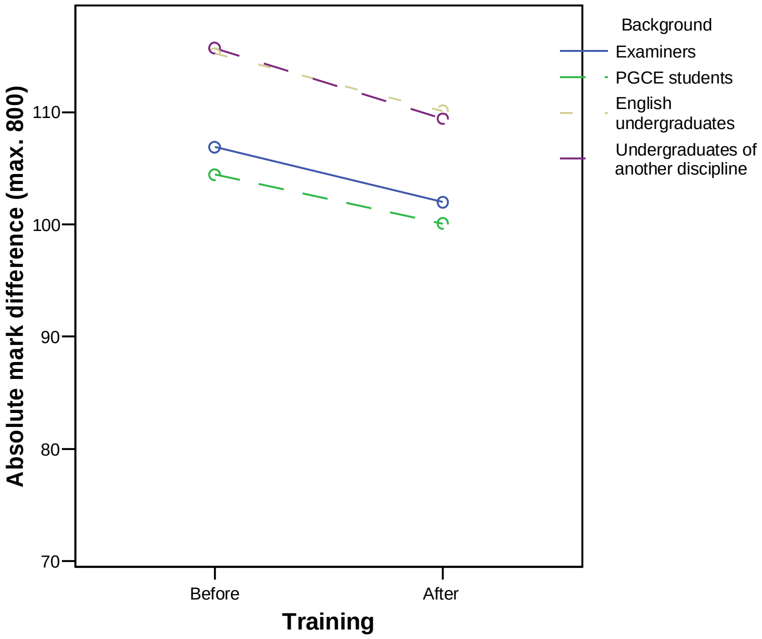
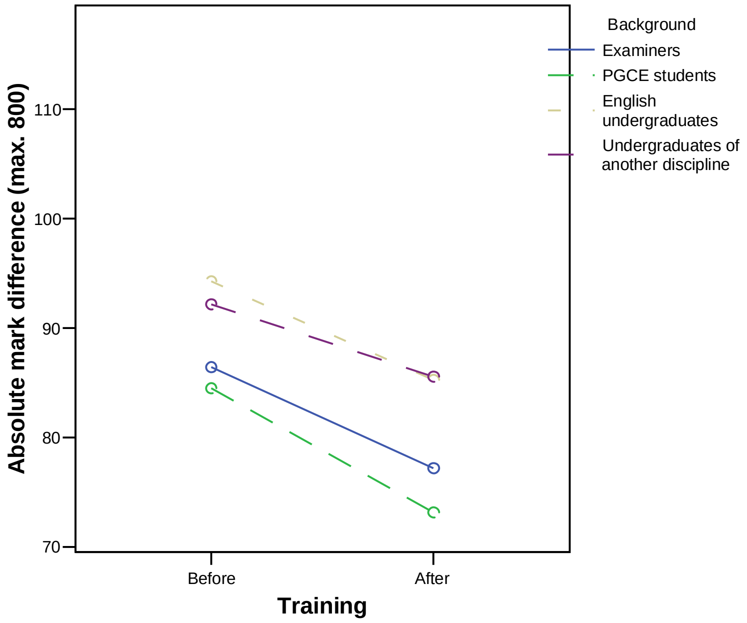


Figure 14 The effect of marker background and training on the absolute difference in the mean mark awarded to item 2a by all the participants and that awarded by individual participants



Since this item was marked out of 8 it was possible to examine the correlation between the participants' marks and the estimated 'true' marks. As before, a Fisher transformation was applied to the correlation data to allow their use as dependent variables in ANOVA (Clark-Carter, 2006).

There was a significant effect of marker background on reliability for both estimates of 'true' mark. Examiners marked significantly more consistently than all the other groups (*including the PGCE students*) and that PGCE students marked more reliably than the undergraduates (see Figures 15 and 16).

Training had no significant impact on the correlation between the all groups of participants' marking and that of the Principal. However, using the consensual estimation of 'true' mark, training significantly improved the correlation for all groups apart from the PGCE students

Figure 15 The effect of marker background and training on the correlation between the Principal Examiner's and participants' rank ordering of candidates' responses to item 2a

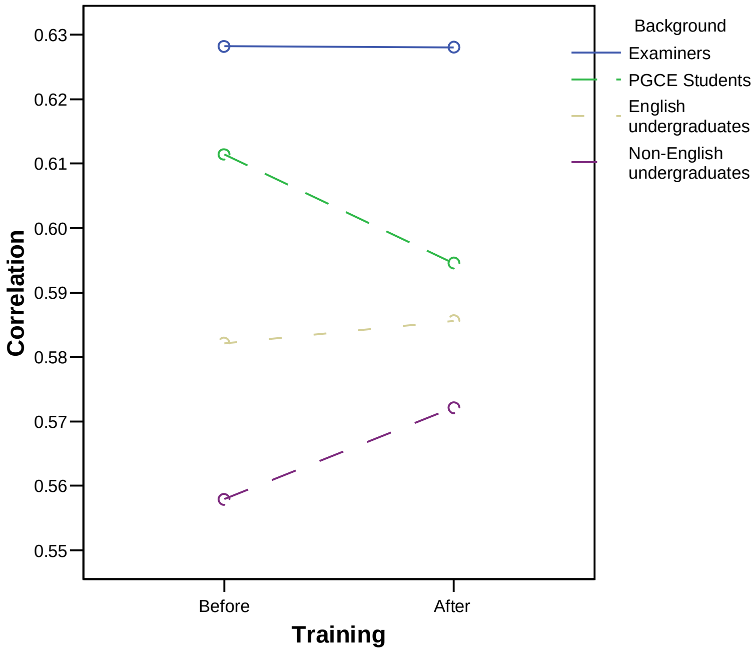
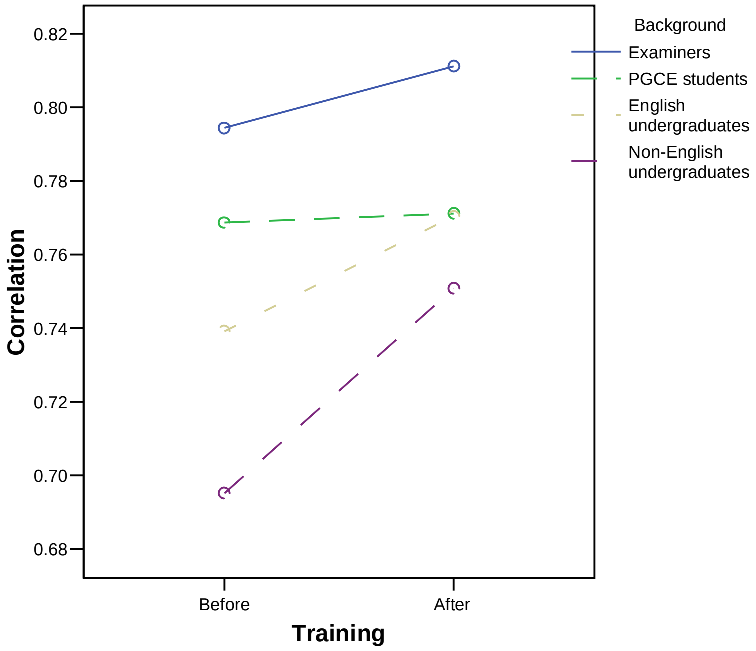


Figure 16 The effect of marker background and training on the correlation between the mean of marks awarded by all participants and the marks awarded by individual participants' to candidates' responses to item 2a



Item 2b (6 marks)

Principal Examiner's mark as the 'true' score - absolute difference in marks awarded to work by the Principal Examiner and the participants

For this item the findings varied according to the definition of 'true' mark employed (see figures 16 and 17). When a hierarchical definition was used there was a significant effect of marker background on the absolute difference in marks awarded. Tukey *post hoc* tests showed that the examiners had lower mark differences than the English undergraduates. However, when a consensual definition was used, there was no significant effect of marker background.

Using a hierarchical definition of 'true' mark there was no significant impact of training on reliability and this was the case no matter what the background of the participants. However, when a consensual definition was used training had a positive effect on reliability for all groups.

Figure 16 The effect of marker background and training on the absolute difference in marks awarded by the Principal Examiner and the participants to item 2b

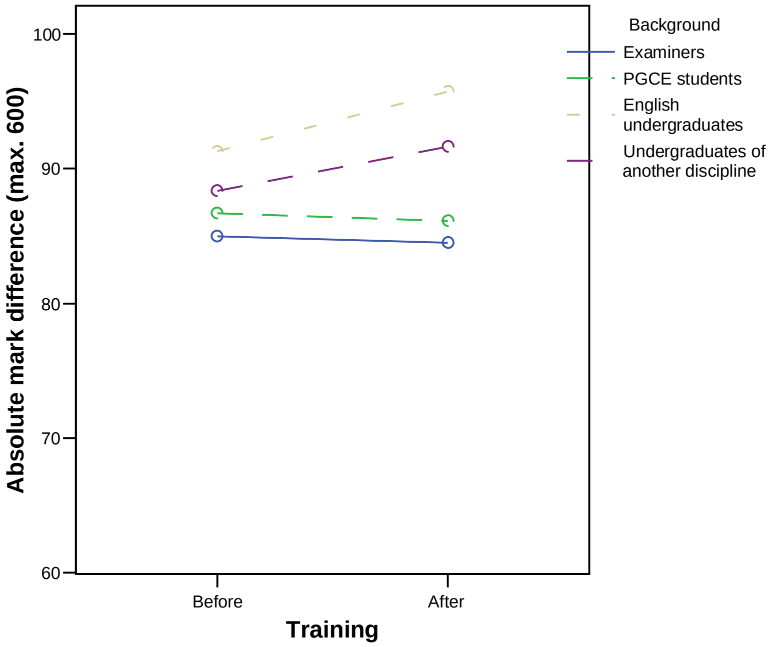
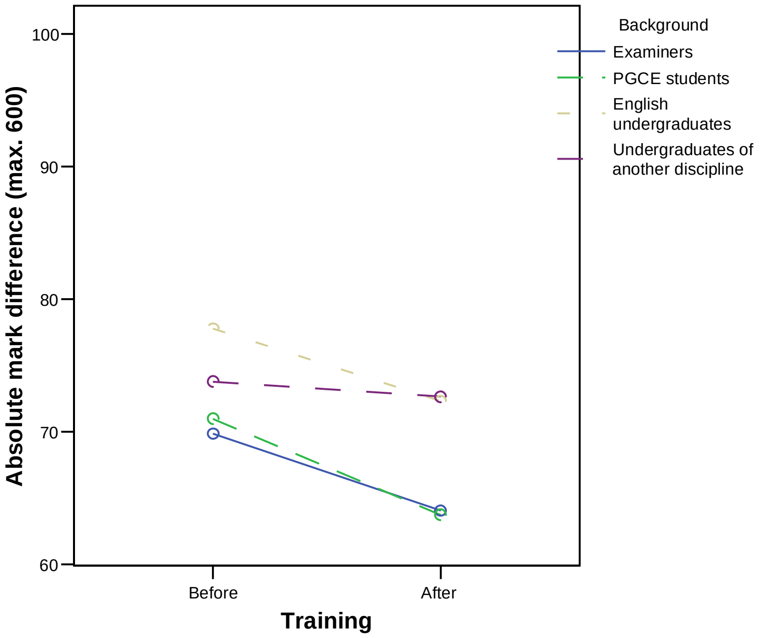


Figure 17 The effect of marker background and training on the absolute difference in the mean mark awarded to item 2b by all the participants and that awarded by individual participants



Discussion

Although some studies have shown that examining and teaching experience tends to be associated with relatively generous marking, other studies have failed to replicate this finding. In this study there was no significant difference in the severity of the marking of the various groups.

There were, however, some differences in the extent to which the groups marked reliably at part-script level. While there was no difference between the groups in the size of the absolute mark differences from either estimation of 'true' mark, they differed in terms of the consistency of their marking. There was no significant difference in the consistency of the Examiners' marking and that of the PGCE students. However, the Examiners' marking was more strongly correlated with both estimates of 'true' mark than that of the undergraduate groups. There were, however, some undergraduates who marked as well as the best examiners. Categorising participants as 'good' markers on the basis of the mean correlation and absolute maker difference of the Examiners revealed that an equal proportion of the Examiners, PGCE students and undergraduates from another discipline fell into the 'good marker' category.

The analyses conducted and conclusions drawn have used the marking reliability of the examiners as a point of comparison (a gold standard). There is some evidence to suggest that the undergraduates did not mark as reliably as the examiners, but that is not to say that they did not mark reliably enough. Equally, it may be that by operational standards the examiners did not mark reliably. Making relative judgements about reliability of marking is unsatisfactory and a technical method of defining an acceptable level of reliability needs to be developed. The conclusions of this study should be reviewed in the light of that definition.

Turning to the reliability with which the individual items were marked, there were no significant differences between the groups of participants in marking items 1a and 1c. These items both required relatively short responses, being marked out of 3 and 6 marks respectively. There were, however, differences for the equally short response items 1b and 2b marked out of 4 and 6 respectively. Using the hierarchical estimation of 'true' mark, the PGCE students marked item 1b more reliably than the English undergraduates and the Examiners marked item 2b more reliably than the English undergraduates did. It would be difficult to predict which items could be marked reliably by those markers without the subject knowledge and teaching experience of the Examiners and PGCE students. The surface characteristics of the items in terms of the length of response they require are not adequate to base this decision on. There is, however, no evidence to suggest that PGCE students could not mark these kinds of items as reliably as Examiners.

Item 2a required the longest response (the maximum mark was 8) and arguably the most expertise in marking. The Examiners and PGCE students were significantly more reliable in their marking of this item in terms of the absolute mark difference from the estimated 'true' mark. There was however, also a significant difference in the consistency of marking of the PGCE students and Examiners. Examiners' marking was more strongly correlated with the estimate of 'true' mark than that of PGCE students and the other groups.

Could individuals with no teaching experience be employed to mark GCSE English? In general, the examiners marked more reliably than the undergraduates or the English undergraduates. It seems that both subject knowledge and some experience of teaching/teacher training are important to marking reliability. The findings do not support the employment of the latter groups of individuals as examiners. While they mostly responded positively to training, the improvement in the reliability of marking was not sufficient; there remained a significant shortfall in the reliability of their marking compared to that of examiners.

Making a recommendation regarding the possibility of employing PGCE students to mark GCSE English is more difficult. There was no evidence to suggest that PGCE students should not be employed to mark short answer questions. There was however, evidence that PGCE students failed to mark longer answer questions as reliably as examiners. Despite concern regarding the ability of PGCE students to mark longer answer questions, there was no significant difference in the reliability of their marking and that of examiners at the level of part-script. Inconsistencies in their marking at item level cancelled out at part-script level. Nonetheless, it would be inappropriate to conclude that PGCE students could be employed to mark whole scripts (as well as short answer questions) since we have evidence that they would not be marking the longer answer questions satisfactorily. This would particularly impinge on the reliability of the grades awarded to those candidates whose total mark was particularly dependent on their responses to the longer answer questions. These findings highlight the usefulness of systems of item level marking which allow items to be marked by the individuals best suited to the task.

At part-script level training reduced the absolute mark difference from either estimation of the ‘true’ mark but was also unexpectedly associated with a reduction in the size of correlation with the hierarchical ‘true’ mark. The compression of the mark distribution is unfortunate since an explicit function of examiners’ training is to stretch the range of marks awarded so as to avoid compression of the final mark distribution and hence of the grade boundaries. Training, however, also had the negative effect of compressing the distribution of marks awarded by participants. An explicit function of examiners’ standardisation training is to stretch the range of marks awarded so as to avoid compression of the final mark distribution and hence of the grade boundaries. Indeed training materials distributed to senior examiners refer to the desirability of encouraging a spread of marks. It is reassuring that there is no evidence of particular problems of a restricted distribution of marks in GCSE English. Nonetheless, the standardisation training will be re-evaluated in the light of these findings.

Training reduced the absolute mark differences from both the hierarchal and consensual ‘true’ marks at part-script and item level with the exception of item 1a. For this item, training had a detrimental effect, increasing the absolute difference from the ‘true’ mark. There was also some evidence to suggest that the PGCE students would benefit from specially tailored training. PGCE students’ qualitative evaluation of the training given did not highlight any specific problems. Indeed, most evaluations were positive although they said that they would have liked more training. Further research is needed to establish the most appropriate training, perhaps through qualitative work canvassing the views of PGCE students and senior examiners, and through quantitative work testing the impact of customised training on the reliability of PGCE students’ marking.

Certain individuals without teaching experience or subject knowledge are able to mark as well as experienced examiners. The difficulty lies in identifying who they may be and training them appropriately. It may be that other measures of individual differences such as psychometric measures of personality will support this process.

References

Baird, J. & Mac, Q. (1999) *How should examiner adjustments be calculated? - A discussion paper*. AEB Research Report, RC13.

Brown, A. (1995) The effect of rater variables in the development of an occupation specific language performance test. *Language Testing*, v12 n1 p1-15.

Clark-Carter, D. (2006) *Quantitative psychological research: A student's handbook.* Hove: Psychology Press.

Cumming, A. (1990) Expertise in evaluating second language compositions. *Language Testing*, v7 p31-51.

Ecclestone, K. (2001) "I know a 2:1 when I see it": Understanding degree standards in programmes franchised to colleges. *Journal of Further & Higher Education*, v25 n4 p301- 313.

Huot, B. (1988) The validity of holistic scoring: A comparison of the talk-aloud protocols of novice and expert holistic raters. Indiana University

Lumley, T. L., Lynch, B.K. & McNamara, T.F. (1994) A new approach to standard setting in language assessment. *Melbourne Papers in Language Testing*, v3 n2 p19-40.

Meyer, L. (2000a) *The ones that got away - development of a safety net to catch lingering doubt examiners*. AQA Research Report, RC50.

Meyer, L. (2000b) *Lingering doubt examiners: results of pilot modelling analyses, summer 2000*: AEB Research Report.

Michael, W. B., Cooper, T., Shaffer, P. & Wallis, E. (1980) A comparison of the reliability and validity of ratings of student performance on essay examinations by professors of English and professors of other disciplines. *Educational & Psychological Measurement*, v40 p183-195.

Myford, C.M., & R. J. Mislevy (1994) *Monitoring and Improving a Portfolio Assessment System*. Princeton, NJ: Educational Testing Service

Pinot de Moira, A. (2003) *Examiner background and the effect on marking reliability*. AQA Research Report, RC218.

Powers, D., & Kubota, M. (1998a) *Qualifying essay readers for an online scoring network (OSN)*. (RR-98-22) Princeton, NJ: Educational Testing Service.

Powers, D., & Kubota, M. (1998b) *Qualifying readers for the online scoring network: scoring argument essays*. (RR-98-28) Princeton, NJ: Educational Testing Service.

Qualifications and Curriculum Authority (QCA) (2007) *Code of practice 2007*. Great Britain: QCA.

Royal-Dawson, L. (2004) *Is teaching experience a necessary condition for markers of Key Stage 3 English?* AQA Research Report, RC261.

Royal-Dawson, L. and Baird, J. (in preparation) *Is teaching experience a necessary condition for markers of Key Stage 3 English?*

Ruth, L., & Murphy, S. (1988) *Designing writing tasks for the assessment of writing*. Norwood, NJ: Ablex Publishing Corp.

Shohamy, E., Gordon, C., & Kramer, R. (1992) The effect of raters' background and training on the reliability of direct writing tests. *Modern Language Journal*, v76 n1 p27-33.

Spearman, C. E. (1904a) 'General intelligence' objectively determined and measured, *American Journal of Psychology*, 5, 201-293.

Spearman, C. E. (1904b) Proof and measurement of association between two things, *American Journal of Psychology*, 15, 72-101.

Spearman, C. E. (1927) The abilities of man, their nature and measurement (New York, Macmillan).

Weigle, S. (1994) *Effects of training on raters of ESL compositions: Quantitative and qualitative approaches*. Unpublished PhD dissertation, University of California, Los Angeles.

Weigle, S. (1999) Investigating Rater/Prompt Interactions in Writing Assessment: Quantitative & Qualitative Approaches. *Assessing Writing*, v6 n2 p145-178.