The effect of marker background and training on the quality of marking in GCSE English

Michelle Meadows[a]* and Lucy Billington[b]

[a]*Assessment & Qualifications Alliance, UK;* [b]*University of Bristol, UK*

Historically, in the UK, marker selection criteria for national examinations have been a matter of custom and practice. Only recently has empirical evidence of the effect of background on marking reliability begun to be gathered. This study attempts to disentangle the effects of marking experience, subject knowledge and teaching experience on marking reliability in GCSE English. GCSE English examiners (97), trainee English teachers (81), English undergraduates (99) and undergraduates from other disciplines (82) marked 199 part-scripts. Overall the groups marked equally accurately. Examiners' marking was more consistent than that of undergraduates but there was no difference in the consistency of examiners and trainee teachers. There were, however, some undergraduates who marked as well as the best examiners. Particular items were more difficult for certain groups to mark reliably. This could not be predicted by surface characteristics such as maximum mark. A sophisticated, evidence based method of allocating items to markers with different levels of expertise is required.

Keywords: marking reliability; English; selection criteria; accuracy; consistency

Introduction

In the UK, the selection of markers for national assessment systems has up until recently been largely a matter of custom and practice. The examination boards responsible for the delivery of high stakes public examinations such as GCSEs and A levels[1] usually require that markers have academic qualifications, typically a relevant degree or equivalent, and at least three terms' teaching experience. However, a proliferation of examining has meant that markers are in short-supply, with many current examiners coming up to retirement (Suto and Nadas 2008) and a deficit of newer teachers in some subjects (Chevalier and Dolton 2004). Moreover electronic marking technology has provided the facility for items within an examination to be marked separately, by individuals with different experience and qualifications. A crude distinction has been drawn between *general markers* and *examiners* (QCA 2009), the former being employed to mark items that do not require subject expertise. Such distinctions ought to be evidence based and research into the relationship between marker background and marking reliability is crucial in determining future recruitment practices and how individual items should be assigned to different marker groups.

A number of studies have attempted to identify factors that might aid the identification of those individuals likely to mark most reliably and those who are likely to require additional training or

[1] GCSEs (General Certificate of Education) and A levels are academic qualifications awarded in specified subjects. GCSEs are generally taken in a number of subjects by students aged 14–16. A levels are usually taken in 3 or 4 subjects by students aged 17-18 and are considered the standard qualifications for assessing the suitability of applicants for university entrance.

Copyright © 2013 AQA Education and its licensors. All rights reserved.

AQA Education (AQA) is a registered charity (number 1073334) and a company limited by guarantee registered in England and Wales (number 3644723). Our registered address is AQA, Devas Street, Manchester M15 6EX.

monitoring. Marking reliability has been defined in a number of ways. Some studies have focused on markers' 'accuracy', sometimes calculated as the mean difference between participants' marks and the 'correct' or 'correct' mark (usually defined as the mark given by the most senior examiner but sometimes defined as the mean of all examiners' marks). This measure gives an indication of the relative severity of marking but over-estimates accuracy as positive and negative differences from the 'correct' mark cancel out. The absolute mean difference, in which all differences are given positive values, gives no indication of the relative generosity of marking but is a more useful measure of accuracy. Recently some researchers have preferred to report frequency of agreement, defined as the overall proportion of perfect agreement between a marker and the 'correct' mark (see Bramley 2007). Some studies have reported measures of marking 'consistency' which is the extent to which examinees' performances are ranked in the same order as that determined by the 'correct' mark. Accuracy and consistency are likely to be correlated - an accurate marker would also be a consistent marker. However, marking may be consistent yet inaccurate, for example a marker might consistently under-reward work by X marks. In systems in which whole scripts are marked, marking consistency may be considered sufficient as adjustments can be applied to compensate for severity/generosity – a common practice in UK awarding bodies. Unfortunately these adjustments are imprecise (Baird and Mac, 1999); ideally marking ought to be accurate. Indeed in systems in which marking occurs at item level quality control measures focus upon accuracy rather than consistency.

Let us now turn to the studies' findings. There is some evidence that compared to experienced markers, inexperienced markers tend to mark more severely and employ different rating strategies (Ruth and Murphy 1988; Huot 1988; Cumming 1990; Shohamy, Gordon and Kraemer 1992; Weigle 1994, 1999), although training may remove these differences (e.g. Weigle 1999). However, not all studies have replicated the relationship between inexperience and marking severity. For example, Meyer (2000a, 2000b), investigating marking in GCSE English Literature and Geography, found that length of experience and a senior examiner's rating of the marker's performance rarely proved useful as predictors of whether an examiner's marks would require adjustment to correct for severity or generosity. Pinot de Moira (2003) studied the relationship between examiner background and marking consistency and accuracy across seven A levels. She found that the composition of an examiner's marking allocation in terms of the kind of school or college the examinees had come from had far more influence on accuracy than accessible aspects of an examiner's background, such as qualifications, present occupation or seniority. The only personal characteristic found to be significant in explaining examiner reliability was the number of years of marking experience. Although as Royal-Dawson (2004) pointed out, this characteristic was confounded because reliable examiners were engaged year after year and poor markers were not.

Studies have also focussed upon whether teaching experience is a necessary requirement for reliable marking. Shohamy, Gordon, and Kraemer (1992) studied inter-marker reliability in the assessment of English as a foreign language (EFL), using markers who were either professional, experienced EFL teachers or lay people (native English speakers). Half were trained in one of the three marking procedures (holistic, analytic and primary trait scoring). Relatively high inter-marker 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 or background, but the overall inter-marker reliability coefficients were higher for trained markers than they were for the untrained ones. Shohamy *et al*

concluded that a focus on training rather than the background of new markers was appropriate.

Brown (1995) investigated the impact of teaching background on the reliability of ratings in an oral test measuring Japanese Language skills of Australian tour guides. Assessors were either from the tourist industry or they were experienced teachers of Japanese as a foreign language. Overall background had no effect on rating severity or 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 rating criteria: teachers were more severe in ratings of grammar, expression, vocabulary and fluency, whereas non-teachers gave more severe 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.

Working in the US, Powers and Kubota (1998a) investigated whether non-teachers could accurately mark essays written by graduate students seeking admission to programmes in business management. They compared the marking quality of experienced and inexperienced markers. The experienced markers had previously marked similar essays, had graduate degrees and taught in relevant university courses. The inexperienced group either did not have graduate degrees or were not currently teaching relevant courses and had no experience of marking essays. Essays were marked before and after training. After training, all markers but especially the inexperienced, improved significantly in their accuracy. However, several of the inexperienced markers were as accurate as the experienced markers even before the training. Powers and Kubota concluded that there was little relation between background and accuracy and that the current recruitment pre-requisites 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' - used to select candidates for graduate programs in management. As in the previous study, the results suggested that inexperienced markers could be trained to mark the essays accurately. They also collected logical reasoning scores for the markers. The results suggested a possible link between logical reasoning and marking accuracy.

In England, Royal-Dawson and Baird (2009) explored whether teaching experience was a necessary recruitment criteria for National Curriculum tests in English taken at age 14. They compared the marking accuracy of four types of markers with an academic background in English but different amounts of teaching experience: English graduates, trainee teachers, teachers with three or more years' teaching experience, and experienced markers. Accuracy was defined in two ways: the absolute difference between the participants' marks and those of the most senior examiner; and the absolute difference between the participants' marks and the mean marks awarded by all participants. This dual approach had the advantage of not assuming perfect marking by the senior examiner and that there is only one valid judgement about the mark that any response is worth. The samples were small (group sizes varied between 9 and 17 markers) which may have implications for the statistical power of analyses and generalisability of findings. Overall, whichever definition of accuracy was used, there was little difference in the accuracy of the different types of marker. There was however evidence to suggest that classroom experience was needed to accurately mark curriculum specific items relating to Shakespeare. Interestingly experience of marking seemed to reduce the accuracy of marking of a reading task. Royal-Dawson and Baird suggested that the task might have varied from that of previous years in which the markers had experience and that there had been some negative transfer of earlier training. They went on to propose a

rudimentary model for the allocation of markers with varying levels of expertise to different item types. The model was based upon the level of detail encompassed by the scoring criteria and the level of curriculum specificity of the items.

Similarly, Suto and Nadas (2008) compared the marking accuracy of expert and graduate markers in GCSE Physics and Mathematics. Experts had experience of both teaching and marking. Graduates had experience of neither but both groups had a relevant degree. Accuracy was defined as the proportion of raw agreement between the participants' marks and those of the most senior examiner, although the study also reported relative and absolute differences between the participants' marks and those of the most senior examiner. There were very few differences in the accuracy of experts and graduates for either subject. The groups significantly differed on just one question (out of twenty) for mathematics and two questions (out of thirteen) for physics. In any case, the differences in accuracy were small. They came to a similar conclusion to that of Baird and Royal-Dawson (2009), that the selection criteria for GCSE Mathematics and Physics examiners could be relaxed. Subject-specific question features were however related to marking accuracy. Questions requiring markers to use more complex 'reflective' thought processes were marked less accurately than those entailing only simple 'intuitive' judgements. Such differences could form the basis of a rationale for assigning particular questions to different marker groups with different levels of expertise.

Suto, Nadas and Bell (2009) conducted one of the most comprehensive investigations into the factors affecting marking accuracy. Their study focussed upon an International Biology examination designed for 16 year-olds. Forty-two markers participated, comprising five groups: experienced examiners, Biology teachers, graduates in Biology, graduates in other subjects, and non-graduates. The design of their study enabled the investigation of the relative effects on marking accuracy of marking experience, teaching experience, highest education in a relevant subject, and highest education in any subject. In addition, they explored three aspects believed to impact upon the demands of the marking task: cognitive marking strategy complexity, of the maximum mark the question is worth, difficulty of the question for the examinee. In general, marking accuracy was high, and all five groups marked questions requiring simple marking strategies extremely accurately. For those questions requiring more complex marking strategies, highest general education, highest relevant education, target grade and total mark were found to be the most important predictors of accuracy. However, with sufficient training the accuracy of some markers with only GCSEs or A levels was found to be comparable to that of many markers with higher qualifications. Having teaching experience and marking experience were significant but less important predictors of accuracy, although highest general education and marking experience were closely associated making it difficult to partial out their effects. Suto *et al* recognised that a key limitation of the study was its reliance on a single definition of marking reliability, the proportion of exact agreement with the marks awarded by the most senior examiner of the assessment. Thus the senior examiner was considered infallible and there was only one valid judgement about the mark that a response is worth. These assumptions are more likely to hold for the points based marking of relatively short answers in say Maths than level of response based marking of longer answers in English for example.

To summarise, research seems to suggest a tenuous relationship between elements of marker background and marking accuracy, thus, supporting the relaxation of recruitment criteria for some assessments in some subjects. Much of this research, however, is based on relatively small sample sizes with consequent effects on the statistical power of the between group comparisons. Studies have rarely used multiple measures of reliability (consistency

and accuracy, for example) and multiple estimations of the 'correct' mark. Furthermore, few studies have explored the relative effects of different marker characteristics on reliability. Studies have varied in the extent to which the effects of subject knowledge, marking and teaching experience have been confounded.

The current study explores the reliability with which individuals with distinctly different backgrounds mark GCSE English to disentangle further the effects of examining experience, teaching experience and subject knowledge. The study attempted to replicate and extend the findings of Royal-Dawson and Baird (2009) with a different assessment of English and much larger samples of markers. Since it is possible that the relationship between markers' background and marking reliability will vary with the kind of item being marked and the cognitive strategy thus employed, participants were required to mark a mixture of items requiring both short and longer responses. This analysis is particularly informative for systems in which marking occurs at item rather than paper level, allowing markers with different characteristics to be recruited to mark different item types.

The study used three measures of marking quality: severity, accuracy and consistency. Accuracy and consistency were assessed against two estimations of the 'correct' mark, a hierarchical and a consensual estimation. The former being the mark given by the most senior examiner of the assessment, 'correct' mark is operationalised by UK awarding bodies in this way. The latter being the mean mark awarded by all the participants. This approach is closer to that embodied by classical test theory (Spearman 1927), in which the 'correct' mark is given by the pooled judgement of an infinite number of markers, although it is usually assumed that all markers are of equal calibre. Taking both approaches aided generalisation to assessment systems employing either approach and guarded against the study's conclusions being influenced by error in the senior examiner's marking.

Method

Four groups of participants marked the same 199 cleaned GCSE English higher tier part-scripts. Part, rather than whole, scripts increased the variety of work marked in the time available. The part-scripts included two questions: the first required two relatively short answers and one slightly longer answer; the second required two longer answers (see Figure 1). The part-scripts were also marked by the Principal Examiner responsible for setting the paper and mark scheme and training markers during the live examination.

Examinees 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)**

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

GCSE English was considered a suitable subject because historically there is evidence of relative unreliability in marking. The question papers include a variety of items possibly requiring different levels of skill, requiring different cognitive processing strategies, and with the potential for valid disagreement over the worth of responses. The subject is not so specialist as to make reliable marking by non-subject specialists impossible.

The groups of participants were selected to enable the relative importance of previous examining experience, subject knowledge and teaching experience to marking reliability to be assessed (see Table 1). A short screening questionnaire ensured that participants had the requisite amount of teaching experience/subject knowledge for inclusion. All participants spoke English as their first language. The experienced markers had taught GCSE English for at least three years and had examined GCSE English in the past year (but had not marked the examination used in this study). The trainee English teachers (students training for a Postgraduate Certificate in Education (PGCE) to teach English in secondary schools and colleges) had first class or upper second undergraduate degrees in English or English Literature. They had limited experience of teaching as part of their teacher training degree but no formal experience or training in marking, though they may have had some general training in assessment as part of their postgraduate studies. Although both the experienced markers and the PGCE students had some teaching experience, they differed considerably in the extent of that experience. Neither group of undergraduates had any formal experience or training in marking or any formal teaching experience. The undergraduates from 'other disciplines' had not studied English or English Literature past GCSE. The groups of undergraduates were recruited from several universities across the northwest of England with varied entry requirements, from the most demanding (three As at A level) to lower demands (typically three Cs at A level).

A weakness of this quasi-experimental design is that it can only go so far in teasing out the relative importance of marking experience, subject knowledge and teaching experience. For example, it was impossible to recruit participants with high levels of marking experience with no subject knowledge. Further, the groups systematically varied in other ways, in general academic ability for instance. The design, however, retains ecological validity in that the groups of participants were those who UK awarding bodies are actively considering recruiting as markers.

Table 1. Groups of markers participating in the study.

	Marking experience	**Subject knowledge**	**Teaching experience**	**N**
Experienced GCSE English markers	high	high	high	97
PGCE English students	low	high	some	81
English/Linguistics undergraduates	none	high	none	99
Undergraduates of another discipline	none	low	none	82

The study was conducted in a marking centre. Initially participants marked 100 part-scripts using the mark scheme. They then received standardisation training which replicated as closely as possible the training used in the live examination. The Principal Examiner explained the mark scheme, item by item. Seven exemplar scripts were chosen to encompass variation in examinees' responses. After participants marked each of these scripts the Principal Examiner discussed the 'standardised' marks with the group. Participants then marked another 99 part-scripts; hence the participants' response to training was assessed.

Scripts were randomly sampled from over 220,000 scripts marked in the summer 2005 examination. Since marking reliability varies with quality of work (Pinot de Moira 2003) care was taken to ensure that the samples covered the full mark distribution. Marking conducted by the Principal Examiner confirmed that the mark distribution of the samples marked before and after training were comparable, having similar means (16.68 and 16.83 respectively out of 27 marks) and standard deviations (3.96 and 3.71 respectively).

Participants were paid for their involvement in the study. It was necessary to pay the examiners at a higher rate than other markers. It was explained that the future employment of examiners would not be linked to their performance. Nonetheless one cannot rule out differential motivation of the participant groups.

Results

Marking quality was assessed by the:

- Relative severity of marking (mean mark given by the marker);
- Marking accuracy (absolute difference between the mark given by the marker and the estimated 'correct' mark);
- Marking consistency (correlation between the mark given by the marker and the estimated 'correct' mark).

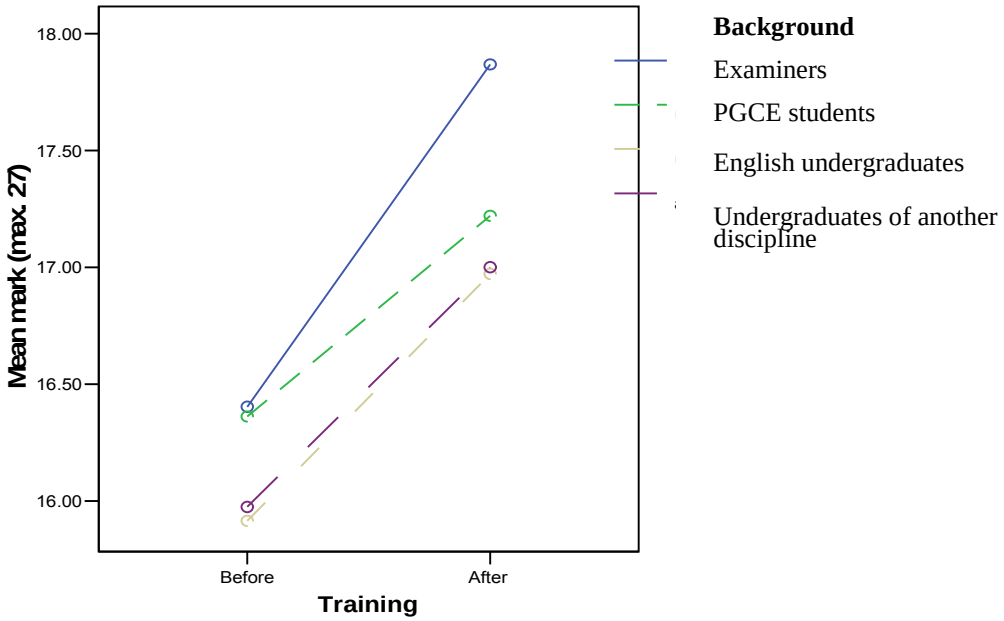
Two of the possible conceptualisations of 'correct mark' (Baird and Meadows 2009) were used:

- Hierarchical: the mark given by the Principal Examiner;
- Consensual: the mean mark awarded by all participants.

Much recent research into marking reliability has taken a multi-faceted Rasch measurement approach to data analysis (Myford and Wolfe 2003, 2004; Schaefer 2008; Wolfe 2009). This is particularly appropriate when the design of the study is incomplete. This study was designed to enable maximum control over the marking situation (all markers marked the same scripts) and so the advantages of a multi-faceted Rasch measurement approach, requiring model fit, were limited. Hence the quality of marking of the groups before and after marker training was investigated using two-way mixed ANOVAs, followed by simple effect analyses and Tukey's *post hoc* contrasts. Unfortunately some participants failed to mark all items on all scripts. This reduced the statistical power of the analyses, particularly those of marking accuracy since only participants with complete marking allocations were included. Estimation of missing marks was inappropriate as it was unlikely that they were randomly spread. Marks for responses that were difficult to mark were most likely to be missing. Had a multi-faceted Rasch approach been taken, this would have violated the assumptions of the model.

Analyses at the level of part-script (27 marks)

Before training, the marking of examiners and PGCE students was on average half a mark more generous than that of the undergraduate groups, although the difference was not statistically significant. After training all groups became more generous to an equal extent (see Figure 2).



Background: F(3, 104) = 2.050, MS = 6.333, p = 0.111, Eta^2 = 0.056, power = 0.512
Training: F(1, 104) = 37.074, MS = 58.103, p < 0.001, Eta^2 = 0.263
Background x training: F(3, 104) = 0.535, MS = 0.838, p = 0.660, Eta^2 = 0.015, power = 0.156

Figure 2. The effect of background and training on the part-script marks.

The range of marks awarded to those part-scripts marked following training was slightly less than those marked before training (see Table 2). Scrutiny of the change in the spread of marks awarded by the Principal Examiner suggests, however, that this was a product of the scripts selected rather an effect of training in encouraging more standardised, or cautious, marking.

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

	Before training			After training		
Background	Mean sd	Min range	Max range	Mean sd	Min range	Max range
Examiners	4.35	14	23	4.09	14	21
PGCE students	3.96	13	22	3.89	14	21
English undergraduates	4.15	14	23	3.97	14	22
Undergraduates of another discipline	3.89	15	22	3.80	13	21
	sd	Range		sd	Range	
Principal Examiner	3.96	20		3.71	18	

Table 3. A summary of the impact of background and training on marking reliability.

Item/ script	Max. mark	Effect of background (df = 3, 353)	Effect of training (df = 1, 353)	Interaction between background & training (df = 3, 353)
Accuracy				
1a	3	$F = 1.203$, $p = 0.312$, $Eta^2 = 0.033$	$F = 1.056$, $p = 0.307$, $Eta^2 = 0.010$	$F = 2.543$, $p = 0.060$, $Eta^2 = 0.067$
1b	4	$F = 3.960$, $p = 0.010$, $Eta^2 = 0.101$	$F = 208.123$, $p < 0.001$, $Eta^2 = 0.663$	$F = 0.802$, $p = 0.496$, $Eta^2 = 0.022$
1c	6	$F = 2.188$, $p = 0.094$, $Eta^2 = 0.058$	$F = 39.660$, $p < 0.001$, $Eta^2 = 0.272$	$F = 0.668$, $p = 0.574$, $Eta^2 = 0.019$
2a	8	$F = 3.806$, $p = 0.012$, $Eta^2 = 0.094$	$F = 7.362$, $p = 0.008$, $Eta^2 = 0.065$	$F = 0.047$, $p = 0.987$, $Eta^2 = 0.001$
2b	6	$F = 2.989$, $p = 0.034$, $Eta^2 = 0.079$	$F = 1.054$, $p = 0.307$, $Eta^2 = 0.010$	$F = 0.688$, $p = 0.561$, $Eta^2 = 0.019$
Part script	27	$F = 2.245$, $p = 0.087$, $Eta^2 = 0.061$	$F = 29.386$, $p < 0.001$, $Eta^2 = 0.220$	$F = 0.645$, $p = 0.588$, $Eta^2 = 0.018$
Consistency				
2a	8	$F = 2.720$, $p = 0.048$, $Eta^2 = 0.071$	$F = 17.584$, $p < 0.001$, $Eta^2 = 0.142$	$F = 0.186$, $p = 0.906$, $Eta^2 = 0.005$
Part script	27	$F = 11.146$, $p < 0.001$, $Eta^2 = 0.087$	$F = 7.364$, $p = 0.007$, $Eta^2 = 0.020$	$F = 5.743$, $p < 0.001$, $Eta^2 = 0.047$

The impact of background and training on marking accuracy and consistency is summarised in Table 3. Almost without exception the findings were the same whether the 'correct' mark was defined consensually or hierarchically. Hence only findings for the latter estimation of correct mark will be reported as this is the standard definition used by UK awarding bodies and that used by the vast majority of previous studies.

Background had no effect on the marking accuracy of part-scripts. On average, the accuracy of the groups varied between 2.6 and 3.1 marks per script, approximately 10 *per cent* of the maximum mark. However, background affected marking consistency,[2] correlations between the marks awarded by the Principal Examiner and those awarded by the participants ranged between 0.64 and 0.70. Examiners and PGCE students marked more consistently than both groups of undergraduates did. There was no difference in the marking consistency of PGCE students and examiners.

Training improved accuracy to the same extent whatever the markers' background but the effect was small, approximately a fifth of a mark per script. However, training had no impact on the marking consistency of either group of undergraduates but appeared to reduce slightly the consistency of the examiners and PGCE students who were more consistent prior to training. Correlations were reduced from 0.70 to 0.68 and 0.69 to 0.66 respectively. Hence, training seemed to both reduce the absolute mark difference from the estimated correct mark and for some groups of participants, reduce the correlation of their marking with the estimated correct mark. One would expect an increase in accuracy to be associated with an increase in consistency. However, this seemingly contradictory finding is explained by the reduced spread of marks post training which impacted on the correlation with the correct mark.

The evidence suggests no difference in the consistency of part-script marking of PGCE students and examiners but the marking consistency of undergraduates, English or otherwise, was poorer than that of examiners. Accuracy of marking did not, however, differ between the groups. Indeed, there were some undergraduates who marked as well as the best examiners. So what constitutes reliable enough marking? Figure 3 shows the correlation and absolute mark difference of the participants labelled by background. The lines represent the mean correlation and absolute mark difference of the examiners. One way of defining a 'good' marker, is one 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 per cent of the examiners, 43 per cent of the PGCE students, 43 per cent of the undergraduates from another discipline and 37 per cent of English undergraduates. So while treating reliability as a continuum suggests an effect of marker background at least on consistency, using a categorical definition might lead to a different conclusion.

[2] A Fisher transformation was applied to correlation data to allow their use as dependent variables (Clark-Carter, 2006)

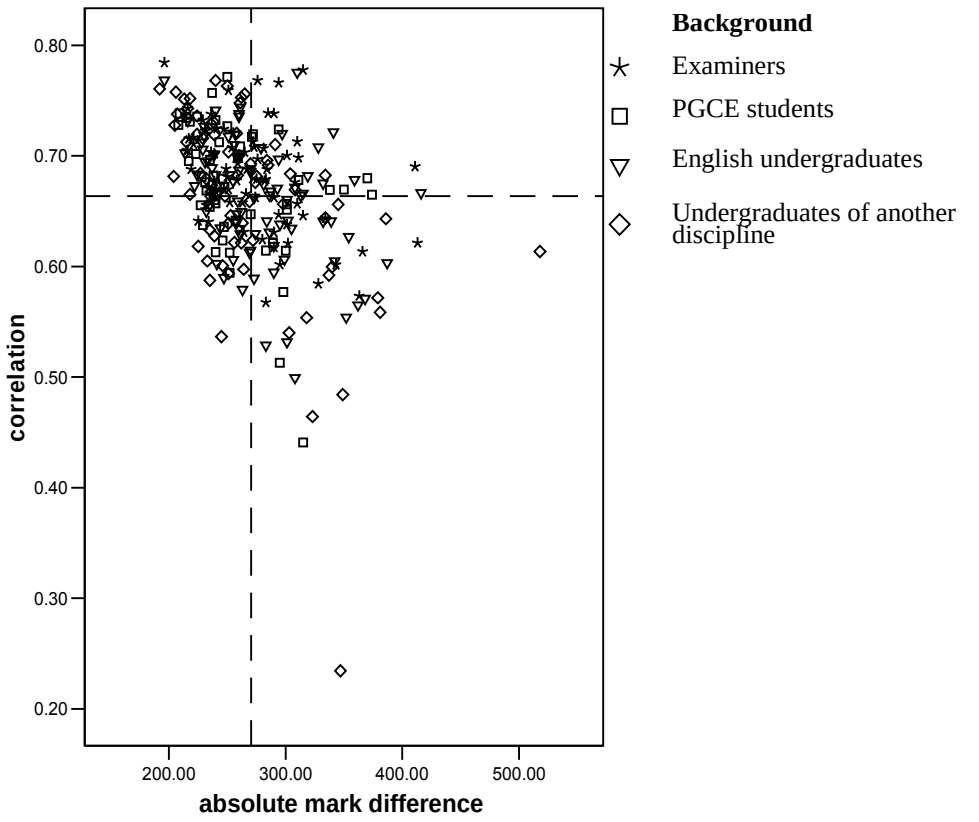


Figure 3. A scatter-plot of the mean absolute difference from the hierarchical true mark against the mean correlation with the hierarchical true mark at part-script level.

Analyses at item level

Item 1a (3 marks): There was no difference in the accuracy with which the groups marked this item. On average, participants varied from the Principal Examiner by 0.4 of a mark. Overall training didn't impact on accuracy although there was a marginal interaction with background. Training decreased the accuracy of the examiners and PGCE students but increased the accuracy of the undergraduate groups. The effect was however extremely small, the largest effect being a decrease in accuracy in the examiners' marking of, on average, 0.06 marks. Nonetheless, training did not have the expected positive impact on marking accuracy for the examiners and PGCE students. While these groups were marking slightly more accurately than the undergraduate groups before training, the training brought their accuracy to a level similar to that of the undergraduates.

Item 1b (4 marks): On average, participants varied from the Principal Examiner by approximately half a mark. PGCE students marked slightly more accurately (by approximately a tenth of a mark) than English undergraduates did, although the difference all but vanished after training. Training increased accuracy equally across the groups, by between 10 and 17 *per cent* of a mark.

Item 1c (6 marks): On average, participants varied from the Principal Examiner by a mark. There was no effect of background on accuracy which increased equally across the groups following training by up to ten *per cent* of a mark.

Item 2a (8 marks): On average, participants varied from the Principal Examiner by slightly more than one mark. There was an effect of background on accuracy. Tukey *post hoc* contrasts were non-significant but the examiners and PGCE students tended to mark more accurately (by approximately a tenth of mark) than the groups of undergraduates. Training improved accuracy to the same extent for all groups (by at best six *per cent* of a mark). The maximum mark for this item (8) allowed marking consistency to be explored. Correlations between the Principal Examiner's marks and those of the participants ranged between 0.56 and 0.63 before training, and between 0.57 and 0.63 after. No matter which estimate of correct mark was used, Examiners marked more consistently than all the other groups (including the PGCE students) and PGCE students marked more consistently than both undergraduate groups. Training didn't significantly affect the marking consistency of any groups. However, using the consensual estimation of correct mark, training improved the consistency of marking overall but had very little impact on the consistency of the PGCE students.

Item 2b (6 marks): On average, participants varied from the Principal Examiner by a little less than one mark. Examiners marked slightly more accurately than English undergraduates (by around 5 *per cent* of a mark). No matter what the background of the participants, training did not improve accuracy.

Discussion

Investigating the effects of marker background using a designed study rather than operational data had advantages including the employment of markers who would not, *a priori*, be considered appropriate and the control of extraneous/confounding variables. However, this limited the study's ecological validity (Chamberlain 2008), for example, the impact of motivation on marking quality is impossible to assess outside live marking. There were also limits to the ecological validity of the training. While every attempt was made to replicate operational examiner training, there was variation. Operationally Team Leaders are used to divide large numbers of markers into smaller groups. This leads to more group discussion and one to one feedback, possibly leading to a more consensual, less hierarchical style. So it is possible that some of the observed effects would not occur in live marking. Moreover, new examiners are routinely given training in addition to marking standardisation. As a minimum this is an overview of the marking process including information on the symbols and terms used. It does not focus on application of the mark scheme. It is possible, but unlikely, that some of the differences between the experienced examiners and other groups would be reduced or even removed by this training.

The study used more than one estimation of the 'correct' mark against which the accuracy and consistency of marking could be measured, that is a hierarchical and a consensual estimation. Reporting focused on analyses using the former estimation as the findings were almost identical. Royal-Dawson and Baird (2009) also reported that analyses based upon these alternative estimations rendered identical conclusions. Error in, or valid disagreement with, the most senior marker's marking was not sufficient to change either studies' conclusions. This increases confidence in the conclusions of previous studies that have used a sole hierarchical estimation (e.g. Suto *et al* 2009).

Although some studies have shown that examining and teaching experience is associated with relatively generous marking (Weigle 1994, 1999 for example), others have failed to replicate this finding (Pinot de Moira 2003 for example). Indeed this study found no difference in the marking severity of the groups. There were, however, some differences in the extent to which the groups marked the part-scripts reliably. There was no difference in the accuracy of the groups but there was a difference in consistency. The examiners' marking was more consistent than that of the undergraduate groups (although there was no difference in the consistency of examiners and PGCE students). There were, however, some undergraduates who marked as well as the best examiners. Categorising participants as 'good' markers on the basis of the mean consistency and accuracy of the examiners revealed that an equal proportion of the examiners, PGCE students and undergraduates from another discipline fell into the 'good marker' category. Only the English undergraduates were under-represented.

Studies of this kind have normally used examiners' marking reliability as a point of comparison (a gold standard). In this study there was 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. Indeed the mean correlation between the examiners' marking and that of the Principal Examiner (0.66) fell short of the correlation reported by Fowles (2006a). She found a correlation of 0.95 between the Principal Examiner's re-marking and the live marking of this question paper. Some of this increased inconsistency in marking may be due to participants marking part-scripts (reliability increases with the length of paper marked as sources of unreliability cancel out); the conditions under which the marking was done; and the examiners' lack of experience of marking this syllabus. Nonetheless, making relative judgements about marking reliability is unsatisfactory. A technical method of defining an acceptable level of reliability would be better. An estimation of the likely number of grade changes associated with a level of reliability seems an obvious basis for this definition. However, Baird and Mac's (1999) analysis of the results of 65 reliability studies conducted by the Associated Examination Board in the early 1980s showed that even with a high correlation of 0.9 between examiners' marking the percentage of grade changes is approximately 50 to 60 *per cent*. According to Baird and Mac, the mean correlation of the examiners in this study would be associated with more than 70 *per cent* grade changes. Proportion of grade changes was not used as an estimate of reliability in this study since part rather than whole scripts were marked. While the grade boundaries could be scaled, this assumes that the items included in the part-script are representative of the whole. However, the part-script represented one section of the paper and tested reading ability whereas the full paper also comprised a second section testing writing ability.

Turning to the reliability with which individual items were marked, the groups were equally accurate in marking items 1a and 1c. These required relatively short responses (marked out of 3 and 6 marks respectively). There were, however, some differences between the groups for the equally short response items 1b and 2b (marked out of 4 and 6 respectively), although the PGCE students marked these items as reliably as the examiners did. Item 2a required the longest response (maximum mark of 8). For this item the examiners' and PGCE students' marking was more accurate than that of the other groups. Further while there was no difference in the accuracy of the examiners and PGCE students, the examiners' marking was more consistent than that of the PGCE students and other groups.

While it has become feasible to allocate items to be marked on an item-by-item basis to diverse groups of markers, there is as yet no established evidence based rationale for assigning questions to different groups. Suto and Nadas (2007) reported that examiners'

marking accuracy is related to a variety of subject-specific question features. For example, in GCSE Mathematics these include: the extent to which alternative answers are possible; the extent to which the question is contextualised; whether follow-through marks are involved; and the Principal Examiner's perception of marking difficulty. To some extent these features differed from those reported for GCSE Physics and it is likely that they would again differ for GCSE English. However, Suto and Nadas (2008) and Suto, Nadas and Bell (2009) went on to further develop the notion of classifying items by the complexity of the thought processes required to mark them.

In this study, surface characteristics of items, length of response for example, were inadequate to predict the level/type of expertise required to ensure marking reliability. However, the perceived need for curriculum specific knowledge in the marking task might be more helpful. Following the study, the subject managers and an opportune sample of four experienced examiners for GCSE English were asked to predict which of the items studied would not require curriculum expertise to mark. They consistently identified items 1a and 1c, that is the items on which there was no difference in the marking reliability of the groups. These findings lend support to Royal-Dawson and Baird's (2009) model for the allocation of markers with varying levels of expertise to different item types based on the complexity of the scoring and the level of curriculum specificity of the items. It seems highly likely, however, that these factors would correlate with the cognitive complexity of marking task. While the development of methods of assessing the cognitive demand of marking is on-going, the senior examining team's judgement of curriculum specificity and cognitive demand may provide practical heuristics for the allocation of items to markers with differing competencies.

So could individuals with little or no subject knowledge and/or teaching experience be employed to mark GCSE English? There was evidence that examiners marked more reliably than the undergraduates and the English undergraduates. Although the design made it impossible to know whether it was subject knowledge and/or some experience of teaching/teacher training that was crucial. However, the differences in marking reliability, while significant, were extremely small. In a subject suffering examiner shortages which have sometimes resulted in examiners marking larger than normal allocations, this small reduction in reliability may need to be offset against the possibly larger impact of examiner fatigue on quality of marking.

Fortunately the employment of undergraduates may not be necessary. There was no evidence to suggest that PGCE students should not be employed to mark short answer questions, particularly if they do not require specific curriculum knowledge, although they failed to mark one longer answer question identified as requiring such knowledge as consistently as examiners. This may, however, have been related to a reduction in the variation in marks awarded following training. Moreover 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 risky to conclude that PGCE students could be employed to mark whole scripts (as well as short answer questions). If particular questions were not marked satisfactorily this would particularly impinge on the reliability of the grades awarded to those examinees whose total mark was particularly dependent on their responses to such 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. Unfortunately GCSE English may be unsuitable for electronic marking which would allow item level marking. Fowles (2006b) found that currently available electronic marking software interfered with the marking process for

longer responses and reduced marking reliability, although software developments may overcome this difficulty.

While there were small but significant effects of marker background on marking reliability, there was large variation in marking reliability within the groups and considerable overlap between them in marking ability. Certain individuals without teaching experience or subject knowledge were 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. As part of this study, participants completed measures of personality and attitude to marking. Analysis of this data is underway and may provide useful tools for marker selection.

Acknowledgements

This study was funded by the National Assessment Agency, disbanded in 2008 but then part of the Qualifications and Curriculum Authority, now the Qualifications and Curriculum Development Authority. The authors would like to thank the Research Committee of the Assessment and Qualifications Alliance and Professor Jo-Anne Baird for comments on the paper.

References

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

Baird, J. and M. Meadows. 2009. *What is the right mark? Respecting markers' views in a community of practice*. Paper presented at the annual conference of the International Association for Educational Assessment, Brisbane, Australia.

Bramley, T. 2007. Quantifying marker agreement: Terminology, statistics and issues. *Research Matters: A Cambridge Assessment Publication* 4: 22-8.

Brown, A. 1995. The effect of rater variables in the development of an occupation specific language performance test. *Language Testing* 12(1): 1-15.

Chamberlain, S. 2008. *Do marking reliability studies have validity?* Paper presented at the annual conference of the International Association for Educational Assessment, Cambridge, England.

Chevalier, A. and P. Dolton. 2004. Teacher shortage: Another impending crisis? *CentrePiece* (magazine of the Centre for Economic Performance, London School of Economics) winter, 14-21.

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* 7: 31-51.

Fowles, D. 2006a. *How reliable is marking in GCSE English?* AQA Research Report, RPA 06 DEF RP 048.

Fowles, D. 2006b. *How well does marking in GCSE English transfer to marking using CMI+ with annotation?* AQA Research Report, RPA 06 DEF RP 047.

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

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.

Myford, C. M., and Wolfe, E.W. 2003. Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement*, 4, 386-422.

Myford, C. M., and Wolfe, E.W. 2004. Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement*, 5, 189-227.

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

Powers, D. and M. Kubota. 1998a. *Qualifying essay readers for an online scoring network (OSN)*. (RR-98-22) Princeton, NJ: Educational Testing Service.
Powers, D. and M. Kubota. 1998b. *Qualifying readers for the online scoring network: scoring argument essays*. (RR-98-28) Princeton, NJ: Educational Testing Service.
Qualifications and Curriculum Authority (QCA). 2009. *GCSE, GCE and AEA Code of practice*. 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 J. Baird. 2009. Is teaching experience necessary for reliable scoring of extended English questions? *Educational Measurement: Issues and Practice* 28: 2-8.
Ruth, L. and S. Murphy. 1988. *Designing writing tasks for the assessment of writing*. Norwood, NJ: Ablex Publishing Corp.
Schaefer, E. 2008. Rater bias patterns in an EFL writing assessment. *Language Testing*, 25(4), 465-493.
Shohamy, E., C. Gordon and R. Kraemer. 1992. The effect of raters' background and training on the reliability of direct writing tests. *Modern Language Journal* 76(1): 27-33.
Spearman, C. E. 1927. *The abilities of man, their nature and measurement* (New York, Macmillan).
Suto, I. and R. Nadas. 2007. The 'Marking Expertise' projects: Empirical investigations of some popular assumptions. *Research Matters* 4: 2-5.
Suto, I. and R. Nadas. 2008. What determines GCSE marking accuracy? An exploration of expertise among maths and physics markers. *Research Papers in Education* 23(4): 477-497.
Suto, I., R. Nadas and J. Bell. 2009. Who should mark what? A study of factors affecting marking accuracy in a biology examination. *Research Papers in Education* 1-31.
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* 6(2): 145-178.
Wolfe, E.W. 2009. Item and rater analysis of constructed response items via the multi-faceted Rasch model. *Journal of Applied Measurement*, 10 (3), 335-347.