NAA ENHANCING THE QUALITY OF MARKING PROJECT: FINAL REPORT FOR RESEARCH ON MARKER SELECTION

Michelle Meadows and Lucy Billington

January 2007

This report was commissioned by the National Assessment Agency

EXECUTIVE SUMMARY

The selection of markers for UK national examination systems is largely a matter of custom and practice. 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 and at least three terms of recent, relevant teaching experience. However, the proliferation of examining and the introduction of computer-based assessment have meant that the need for empirically supported examiner recruitment and selection practices has taken on new importance.

This study explored the value of using measures of personality and attitude as predictors of marking reliability for participants from distinctly different education, teaching and examining backgrounds. Four groups of participants marked the same 199 GCSE English part-scripts that included questions that required short and longer answers. The groups were as follows: 97 experienced GCSE English examiners (high subject knowledge and teaching experience); 81 PGCE English undergraduates (high subject knowledge and some teaching experience); 99 English undergraduates (high subject knowledge and no teaching experience); and 82 non-English undergraduates (low subject knowledge and no teaching experience). The purpose of this design was to help disentangle the association between subject knowledge and teaching experience and the reliability of marking of different item types.

Initially participants marked a batch of 100 part-scripts by applying the mark scheme (no standardisation training had been received). They then received the current marker training for GCSE English and then marked another batch of 99 part-scripts. Participants completed the NEO-FFI that measures five personality domains: neuroticism, extroversion, openness to experience, agreeableness and conscientiousness. At the end of marking, participants completed a measure of their attitude to marking. Factor analysis of this questionnaire gave rise to three factors relating to their enjoyment of marking, their view of the role of judgement as opposed to strict adherence to the mark scheme in marking and their belief that only teachers should be employed as examiners.

Different operationalisations of marking reliability were used: the absolute difference between the mark given by the participant and the estimated ‘true’ mark, and the correlation between the mark given by the participant and the estimated ‘true’ mark. ‘True’ mark was defined both hierarchically and consensually. The effects of training and participants’ background on reliability were explored using analyses of variance. Stepwise multiple regressions investigated the extent to which background, age, gender, personality and attitude were significant independent predictors of marking reliability before and after training.

Results varied according to the operationalisation of marking reliability, the definition of ‘true’ score and the item/part-script being marked. Nonetheless, it was possible to draw general conclusions. The examiners marked more reliably than both groups of undergraduates. Both subject knowledge and some experience of teaching seemed important to marking reliability. Findings did not support the employment of individuals from these groups as examiners.

There was no evidence to suggest that PGCE students should not be employed to mark short answer questions, but they failed to mark longer answer questions as reliably as examiners. Prior to training, there was little difference in the marking reliability of examiners and PGCE students. Unfortunately, the marker standardisation training either failed to improve the reliability of the PGCE students’ marking or even caused it to deteriorate. If PGCE students were to be

employed as markers of longer answer questions they would require customised training. Further research is needed to establish the form of that training.

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 is concluded that it would be inappropriate to employ PGCE students to mark whole scripts without customised training since evidence suggests they would not be marking the longer answer questions satisfactorily. These findings highlight the usefulness of systems of item level marking which allow the marking of items by the individuals best suited to the task.

There was some evidence that older participants tended to mark certain items more reliably than younger participants did. What it was about older participants, over and above their personality, attitude to marking, and marker background that led them to mark certain items more reliably was unclear. Moreover, extremely robust evidence of this age effect would be needed to support the active recruitment of older rather than younger examiners. Equally, it is not immediately apparent why male participants marked some items more reliably than females and vice versa. Again, the evidence is not strong enough to support any discrimination based on gender, which would of course be difficult to defend to stakeholders.

Regarding the use of psychometric measures of personality to predict those individuals likely to mark most reliably, only agreeableness and conscientiousness were positively associated with marking reliability following training. The relationships were relatively weak, accounting for only small amounts of variation in marking reliability, but the difficulties and past failure of previous research in identifying any variables that consistently predict marking reliability (an inherently noisy variable) must be borne in mind. Before trialling the operational usefulness of these measures, the relationship between agreeableness and conscientiousness and marking reliability should be replicated.

Any attempt to use measures of attitude in examiner recruitment and selection would be flawed since applicants would be able to ‘fake good’. Moreover, participants’ attitudes to marking predicted marking reliability prior to training but not following training. Training eradicated the impact of attitudes on marking reliability, surely a positive effect.

Training, however, also had the negative effect of compressing the distribution of marks awarded by participants, despite one of its functions being to stretch the range of marks awarded so as to avoid compression of the final mark distribution and hence of the grade boundaries. It is reassuring that there is no evidence of particular problems of a restricted distribution of marks in GCSE English. Nonetheless, standardisation training ought to be re-evaluated in the light of these findings.

The relationships between marker background, personality, attitude and demographic factors and marking reliability were complex. We need to understand more about the characteristics of items that mediate these relationships before we will be able to predict who will be able to mark particular items reliably. Surface characteristics such as the extent to which expert subject knowledge is required to mark the item do not seem to explain the links between these factors and reliability. Moreover, it may be that the way in which the marker standardisation training was delivered accounts for some of these relationships. The ephemeral nature of the training makes it difficult to know but this possibility will be investigated through discussion with the Principal Examiner.

CAN WE PREDICT WHO WILL BE A RELIABLE MARKER?

Introduction

Background

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). They should have at least three terms' teaching experience which should be recent (usually within the last three years depending on length of experience) and relevant (usually in schools or colleges, but may include university lecturing experience, teaching abroad or private tutoring). Experience of teaching AQA specifications is considered helpful, but not essential.

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, ACCAC, CCEA, 2005) demands that examiners must have relevant experience in the subject but does not explicitly discuss the nature of this experience.

In the UK, 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 awarding bodies to predict 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; Shohmy, 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

[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.

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.

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 or 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? Are there stable or relatively stable individual factors that influence the reliability of marking?

Examiner traits and marking performance

There have been some attempts to link personality traits with marking performance. Branthwaite, Trueman and Berrisford (1981) examined the relationship between 15 markers' scores on the Eysenck Personality Questionnaire and the marks they awarded to an essay. The marks given were unrelated to extroversion, neuroticism or psychoticism scores but were positively correlated with scores on the lie scale. This was interpreted as suggesting that marking may be influenced by desire for social acceptance. That depending on the personality of the marker, considerations of social interaction may bias marker's objectivity. If this were the case then one explanation for low reliability in marking would be the differential desire among markers to appear socially acceptable. Participants in this study marked only one essay in the Higher Education context; it seems likely that the desire for social acceptance would have less influence on the marking of examiners of GCSE and A level scripts, who mark hundreds of scripts of unknown candidates.

Pal (1986) compared the Meenakshi Personality Inventory scores of two groups of four examiners labelled as efficient and inefficient on the basis of the reliability with which they had marked twenty scripts of high school students in the subject of Hindi. Compared with inefficient examiners, efficient examiners had high needs for achievement and dominance, but low needs for affiliation. The two groups of examiners did not significantly vary in their need for exhibition, nurturance, succourance (to have one's needs satisfied by someone or something), abasement, autonomy, endurance or aggression. Given the likely strength of the relationship between personality and the noisy marking reliability variable, it is surprising that Pal found a significant difference between the groups of examiners with such a small sample size.

The small-scale nature of these studies and the sometimes rather ambiguous personality measures used, preclude sensible interpretation of the effect that examiner characteristics can exert on marking reliability. Using a larger sample Greatorex and Bell (2002a and b) had examiners of GCSE English (104), Food Technologies (53) and History (35) complete the Bem Sex Role Inventory. This provides a measure of self-reported possession of socially desirable, stereotypically masculine and feminine personality traits. Examiners who rated themselves highly on the masculinity scales were more likely to be Team Leaders. The masculinity scales are made up of dominant/assertive traits and self-sufficiency/decisive traits. Greatorex and Bell saw this as unsurprising since Team Leaders need to be decisive. The appointment of Team Leaders is under the control of awarding body staff, who presumably perceive these traits to be important in fulfilling the Team Leader role. Team Leaders did not however rate themselves highly on traits that could be useful for developing people skills, which is another important aspect of the role.

Given the association between examiner rank and self-perceived sex-role, investigation of the relationship between examiners' responses to the Bem Sex Role Inventory and marking

reliability may be valuable. However, evidence of no relationship between examiner rank and marking reliability (Pinot de Moira, 2003) makes such an association less likely.

In summary, there have been few studies of the relationship between examiner traits and marking performance. No methodologically robust study has directly investigated this association. Further, it seems likely that the background of an examiner will interact with the type of item being marked to affect marking performance. Indeed this is the basis upon which the marking of certain items by 'clerical' markers has gone ahead in the UK. The National Foundation for Educational Research (NFER) conducted an online marking pilot for Year 7 Progress Tests in mathematics and English. They considered, among other issues, the effect of using unskilled and semi-skilled examiners to mark specifically chosen items (Whetton and Newton, 2002). The marks arising from the unskilled and semi-skilled examiners, once adjudicated by supervisors, were very similar to those arising from expert markers. A similar, though less extensive, pilot study was undertaken by AQA in the marking of GCE Chemistry (Fowles, 2002). The focus of the study was the reliability of e-marking in comparison with conventional marking. The results suggested that, with carefully chosen items, clerical marking could provide a reliable alternative to the use of experienced examiners. Given the findings of the research reviewed, it is questionable whether there would have been differences in the marking reliability of these groups of markers if more demanding items had been included.

The use of psychometric measures of personality in employee selection

Personality tests are widely used within organisational settings for the purpose of personnel selection (Levy-Leboyer, 1994; Anderson and Cunningham-Snell, 2000; Buchanan and Huczynski 2004). The California Personality Inventory, Eysenck Personality Inventory, Guilford-Zimmerman Temperament Survey, and Myers Briggs Personality Type Indicator are among some of the most well known instruments for personality assessment (Salgado, 1997). The underlying rationale for the use of such tests is the notion that personality dimensions are predictive of job performance and future career success.

The relationship between personality and job performance has been of much interest to researchers working in the field of industrial-organisational psychology over the past century or so. Barrick, Mount and Judge (2001) argue that this research can be categorised into two distinct phases. The first phase comprises studies conducted from the early 1900s to the mid 1980s, and is characterised by primary studies which investigated relationships between individual scales from multiple personality inventories and various aspects of job performance. The overall conclusion of this body of research was that personality and job performance were unrelated. Some commentators have sarcastically referred to this as the time when we had no personalities. The second phase spans the period from the mid-1980s to the present and is characterised by the use of the five factor model, or some variant, to classify personality scales. The use of meta-analytic methods to summarise results quantitatively across studies is another key feature. The findings of primary and meta-analytic studies using the five factor model constructs suggest that *"in contrast to the previous era…we do have a personality and that at least some aspects of it are meaningfully related to performance"* (Barrick *et al,* p. 10).

The five factor model arose from systematic efforts to organise the taxonomy of personality. Costa and McRae are the most influential advocates of this approach and have demonstrated that the five factors accounted for the majority of the variance in both self-rating and personality inventory responses, based upon either self-ratings or ratings by persons who knew the targeted individuals well (McCrae and Costa, 1987; Costa and McCrae, 1992a). There is some

disagreement over the names and content of the five factors. They are, however, generally labelled as follows: (1) emotional stability (calm, secure, and non-anxious), or conversely, neuroticism; (2) extroversion (sociable, talkative, assertive, ambitious, and active); (3) openness to experience (imaginative, artistically sensitive, and intellectual); (4) agreeableness (good-natured, cooperative, and trusting); and, (5) conscientiousness (responsible, dependable, organised, persistent and achievement orientated) (Goodstein and Lanyon, 1999). The factors appear in this order (NEOAC) in the Costa and McCrae (1992b) NEO Personality Inventory – Revised, which is currently the gold standard measurement instrument.

Barrick and Mount (1991) were among the first to introduce this personality framework to the industrial-organisational psychology field. They conducted a meta-analysis of 117 research studies that reported statistical relationships between measures of at least one of the five factors and actual job performance. They distinguished three kinds of job performance measures: job proficiency measures, such as productivity indices and performance ratings; training proficiency measures, such as the number and quality of post-training work samples and length of time to complete training; and personnel data, such as salary level, length of service, and number of promotions. They also differentiated five occupational groups: professionals, police, managers, sales, and skilled/semi-skilled. These performance criteria and occupational groups allowed Barrick and Mount to examine whether the five factors could predict job success equally well across occupations and performance levels, and regardless of how performance was measured.

Not surprisingly, conscientiousness emerged as the most consistent predictor of job performance. This was true for all occupational groups. This aspect of personality seems to tap traits important to the successful completion of tasks in all job types. That is, those individuals who possess traits associated with a strong sense of purpose, obligation, and persistence generally perform better than those who do not. Similar findings have been reported in educational settings where correlations between conscientiousness and educational achievement (Digman and Takemoto-Chock, 1981; Smith, 1967) and vocational achievement (Takemoto, 1979) have consistently been reported in the region of 0.50 to 0.60. Extraversion was also a valid predictor for two out of the five occupations - managers and sales. Barrick and Mount observed that for both types of job, interaction with others forms a significant part of the job role. Thus, it seems intuitive that being sociable, gregarious, talkative, assertive, and active would lead to effective performance in these jobs, rather than others.

Both openness to experience and extraversion predicted success in training for all occupations. Since extraversion assesses traits associated with general activity level (talkative, active, assertive) and sociability this relationship is to be expected. Barrick and Mount argued that individuals who score high on the openness to experience dimension (intelligent, curious, broad-minded, and cultured) may respond well to training because they are more inclined to have positive attitudes towards learning in general. A number of researchers have shown that a key component in the success of training programs is the attitude of the individual when s/he enters the training program. Sanders and Vanouzas (1983), for example, demonstrated that the attitude and expectations of trainees influences whether or not learning occurs. Thus, according to Barrick and Mount measures of openness to experience may help to identify individuals that are ‘training ready’ and, consequently those who would benefit most from training programs. Furthermore, openness to experience has the highest correlation of any of the personality dimensions with measures of cognitive ability (McCrae and Costa, 1987). It may be that openness to experience is actually measuring ability to learn as well as motivation to learn.

The correlations for emotional stability (or neuroticism) were relatively low. One possible explanation for this finding may be that individuals with serious problems of emotional stability are absent in the workplace, being either self-selected out or unable to work regularly (Goodstein and Lanyon, 1999). Interestingly the coefficient for professionals on this dimension was negative, suggesting that individuals who are worrying, nervous, emotional, and high-strung are better performers in these jobs. Agreeableness did not seem to be an important predictor of job performance, even in those jobs involving a significant social aspect (sales or management, for example). This finding is in direct opposition to the other socially based personality dimension, extraversion. "*Thus, it appears that being courteous, trusting-straightforward and soft-hearted has a smaller impact on job performance than being talkative, active and assertive"* (Barrick and Mount, p. 21).

In another meta-analysis of studies that had reported a positive relationship between the five factors and job performance, Tett, Jackson and Rothstein (1991) found all personality dimensions were valid predictors of job performance. In contrast to Barrick and Mount's study, agreeableness was the strongest predictor of job performance, followed by openness to experience, emotional stability, conscientiousness, and extraversion. Goldberg (1993) has described the differences in findings based on a similar body of knowledge as 'befuddling'. There are, however, a number of reasons why Tett *et al* and Barrick and Mount arrived at different conclusions. Salgado (1997) attributes the differences to the fact that Tett *et al* only used confirmatory studies (those based on hypothesis testing or on personality-orientated job analysis) in their analysis. Alternatively, Goodstein and Lanyon (1999, p. 296) argue that the *"the differences in the strength of the relationships in the two studies are presumably due to the different jobs that were involved in the two studies."* Furthermore, whilst the specific pattern of results from the two studies differs, they both strongly confirm the utility of using the five personality factors as predictors of on-the-job performance.

The findings of more recent studies are most consistent with those of Barrick and Mount. In his meta-analytic research, Salgado found that conscientiousness and emotional stability were valid predictors across job criteria and occupational groups. Extraversion was a predictor for two occupations and openness and agreeableness were valid predictors of training proficiency. Furthermore, Salgado's research included studies conducted in the European Community, whilst previous studies included only studies conducted in the United States and Canada. Thus, it seems personality measures can predict job performance across different countries and cultures. At the turn of the century, Barrick, Mount and Judge (2001) conducted a meta-analysis of 15 prior meta-analytic studies that have investigated the relationship between the five factors and job performance. Results largely mirrored those of Barrick and Mount's earlier study. They called for a moratorium on meta-analytic studies, and suggested that researchers embark upon a new research agenda with the aim of further enhancing understanding of personality-performance linkages.

Personality measures, especially those based on the five factor model, are valid predictors of job performance. Until 1991, only two personality inventories had been developed within the five factor framework. However, today there are over 15 inventories in the USA and Europe developed within this framework and used in organisational settings. Research comparing the criterion validity of the personality dimensions when assessed using five factor model-based inventories and non-five factor model based inventories confirms that practitioners should use the former to make personnel selection decisions (Salgado, 2003).

Current Study

This study explores the utility of psychometric measures of personality, specifically a five factor model-based inventory, in marker recruitment for individuals with different education, teaching and examining backgrounds. This research may support the selection and employment of individuals with non-teaching backgrounds as examiners in subjects where there is an examiner shortage. Further, the reliability with which individuals with different education, teaching and examining backgrounds mark different types of item, 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.

It is likely that markers' motivation and attitudes to marking will affect marking quality and these will vary with background. It is possible, for example, that markers from non-teaching backgrounds will be less motivated to mark candidates' work accurately. Hence, a questionnaire was constructed to measure attitude and motivation (although it is impossible to test fully the effect of motivation on the quality of marking in a non-live setting).

The extent to which measures of personality, motivation and attitude prove useful as predictors of marking reliability may interact with marker background variables. They may be more useful in predicting the reliability of marking of new examiners rather than experienced examiners, or of examiners from non-teaching backgrounds. The investigation was therefore conducted with participants from distinctly different education, teaching and examining backgrounds. This also provided an opportunity to attempt to replicate the findings of a previous AQA study: 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).

It is likely that the relationship between markers' personality, motivation and attitude, and marking reliability will interact not only with their background, but with the kind of item being marked (as was clearly demonstrated in the study of Key Stage 3 English marking). For example, a highly motivated, able, conscientious 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.

Participants initially marked by simply following the mark scheme, that is, with no formal marker standardisation training. They were then trained and required to continue marking. A measure of responsiveness to training was thus generated. The relationship between markers' personality, motivation and attitude and their responsiveness to training could therefore be examined. Whether any of the inexperienced markers were as accurate as the experienced markers even before the training, as found Powers and Kubota (1998a), could also be investigated.

In summary, the study attempts to address the following kinds of question:

- How, and by how much, can the quality of marking be improved by knowing about easily collected marker characteristics?
- What kind of background is necessary to enable an individual to mark reliably?
- What level of education, subject knowledge and teaching experience is needed?
- Does this vary according to the kind of item being marked?
- Can some kinds of item be marked reliably by anyone, regardless of background?

- How important is the attitude and motivation of individuals from different backgrounds to reliable marking?
- To what extent can psychometric measures of personality predict marking reliability?
- Does this vary with an individuals' background and the type of item they are marking?
- Are some markers more responsive to training than others are?
- Does this vary with background, personality, attitude and motivation?

In other words, the study was designed to inform: the criteria used to select examiners; the kinds of items individuals with different traits, abilities and backgrounds are best able to mark reliably; and the kind of support and training that would enable them to mark reliably.

Methodology

Four groups of participants were recruited to mark the same two hundred 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. Copies of the question paper and mark scheme are in Appendix 1.

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 methods 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 groups of participants are described in Table 1. A short screening questionnaire ensured that participants had the requisite amount of teaching experience and subject knowledge to qualify for inclusion. For example, participants in the English or non-English undergraduate group had negligible or no teaching experience.

Table 1 Groups of markers participating in the study

	subject knowledge	teaching experience	N
Experienced GCSE English A Paper 2 markers	high	high	97
PGCE English undergraduates	high	some	81
English/Linguistic undergraduates	high	none	99
Non-English undergraduates	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 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
Completed NEO-FFI

Day 3:
Marked another batch of 100 GCSE English A paper 1H part scripts
Completed marker feedback, attitude and motivation questionnaire

Participants completed a condensed version of the NEO-PI (240 items), known as the NEO-FFI (60 items). The shorter version was deemed more appropriate for these research purposes, taking less time to complete, but providing a comparable amount of information. Of the 60 items, 12 related to each of the five personality domains. Participants received scores reflecting their level of neuroticism, extroversion, openness to experience, agreeableness and conscientiousness. Each item consisted of a statement, for example 'I am not a worrier', for which participants were required to choose one of five responses; strongly disagree, disagree, neutral, agree or strongly agree.

At the end of marking, participants were canvassed on their overall experience of marking and completed a 44 item bespoke measure of attitude and motivation (Appendix 2). The questionnaire measured participants' enjoyment of marking, the extent to which they believed anyone given training can mark, the level of care they believe should be applied to marking, an evaluation of their own marking abilities, and the role of judgement versus strict adherence to the mark scheme. Each item consisted of a statement, for example 'Marking candidates' work is a rewarding experience', for which participants were required to choose one of five responses; strongly disagree, disagree, neutral, agree or strongly agree. Participants were however, aware of the purpose of the study and the possibility of motivated responding must be kept in mind.

Results

Initial analysis of the NEO-FFI

For all questions missing data were treated as a neutral response (as advised by the NEO-FFI scoring manual (Costa and McCrae, 1992c). No participant had more than two missing items in total or more than one missing item for a particular domain, meaning enough data were present to include all participants in the analysis. Scores for each of the five personality domains were calculated by summing the responses from the relevant 12 questions for each domain, so total domain scores could range from 0 (low) to 48 (high). Reliability checks revealed high internal consistency for each scale. Cronbach's alpha values ranged from 0.72 to 0.88 which were in the acceptable range and consistent with typical values obtained by the scale authors (0.68 to 0.86, for example).

Initial analysis of the motivation and attitude to marking questionnaire

Principal components analyses of the 44 attitude and motivation items examined the internal structure of the survey. An initial oblimin rotation gave rise to three factors that were not correlated. Hence, a varimax rotation was applied. Examination of the scree plot suggested the retention of three factors with eigenvalues of 7.06, 4.62 and 2.39. The three sets of items with factor loadings greater than 0.45 were then interpreted. The outcome of this interpretation is summarised in Table 2. For each factor, the three items with the largest factor loadings are reported.

Table 2. Summary of factor analysis of the attitude to marking questionnaire

Factor 1: Enjoyment of marking 11 items with factor loadings >0.45	
I enjoy the experience of marking	0.78
I got a great deal of satisfaction from marking candidates' work	0.78
Marking candidates' work is a rewarding experience	0.75
Factor 2: Only teachers should mark 7 items with factor loadings >0.45	
Awarding bodies should only employ teachers to mark	0.84
Only experienced teachers should be allowed to mark candidates' work	0.80
With the right training, anyone educated to degree level could accurately mark	-0.78
Factor 3: Role of judgement 5 items with factor loadings >0.45	
It is essential to use judgement in awarding marks rather than blindly following the mark scheme	0.57
The mark scheme is a guide to help the marker, it should not be rigidly adhered to	0.55
Rigidly following the mark scheme can mean that some candidates' work is under or over rewarded	0.53

Relationship between attitude to marking and marker personality, background, age and gender

Stepwise regression analyses investigated the independent predictors of scores on the attitude to marking factors (see Table 3). These showed that conscientious participants and examiners tended to enjoy marking most. Perhaps unsurprisingly, examiners and PGCE students tended

to believe that teaching experience is necessary to mark accurately. Participants with low scores on the agreeableness scale were also likely to take this view. It is likely that participants believed that the purpose of the research was to widen the pool from which examiners will be recruited and participants who were more agreeable were less likely to disagree with this stance. This is interesting evidence of the way in which motivated responding can influence questionnaire findings.

Participants with relatively high scores on the neuroticism scale, younger participants and those with high scores on the openness to experience scale tended to believe it is important to use judgement when applying the mark scheme. It is unclear why younger participants were more likely to take this view since the association between youth and this belief is independent of participant background or personality.

Table 3 Independent predictors of the extent to which participants reported enjoying marking

Variable	Beta	t	p
The extent to which participants reported enjoying marking **R^2=0.091, F(2,319)=16.019, p<0.001**			
Conscientiousness	**0.237**	**4.331**	**<0.001**
Examiners	**0.143**	**2.616**	**0.009**
Belief that only teachers should be employed to mark **R^2=0.506, F(3,318)=108.377, p<0.001**			
Examiners	**0.761**	**17.795**	**<0.001**
PGCE Students	**0.150**	**3.551**	**<0.001**
Agreeableness	**-0.126**	**-3.044**	**0.003**
Role of judgement in applying the mark scheme **R^2=0.138, F(3,318)=16.930, p<0.001**			
Neuroticism	**0.235**	**4.446**	**<0.001**
Age	**-0.247**	**-4.553**	**<0.001**
Openness	**0.143**	**2.666**	**0.008**

Analyses of the reliability of marking

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

- 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 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 examiners. 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.

These approaches were used to investigate the quality of marking at both item and script level, before and after marker standardisation training. 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 Examiners' marking of the scripts.

How reliably did participants mark following training?

There was great variation in the extent to which the participants' rank ordering of candidates' work was similar to that of the Principal Examiner (see Table 4). The correlation for one participant's marking of item 1c, for example, was as low as 0.07. No participants' marking was very highly correlated with that of the Principal. The best achieved was 0.84 on item 1b. On average, participant's rank ordering of candidates' work most closely resembled that of the Principal Examiner at the level of part-script (0.66). Higher reliability estimates at script rather than item level are a consequence of inconsistencies at item level cancelling each other out. The marking of item 1c had the lowest mean correlation. The standard deviations indicate comparable variation in the correlation between the Principal Examiner's mark and the participants' marks for each item and the part-script total.

Table 4 Descriptive statistics for the correlation between the marks awarded by the Principal Examiner and those awarded by the participants

Item	Maximum mark possible	Minimum	Maximum	Mean	Std. Deviation
1a	3	0.08	0.72	0.52	0.10
1b	4	0.28	0.84	0.59	0.07
1c	6	0.07	0.63	0.45	0.08
2a	8	0.08	0.76	0.60	0.08
2b	6	0.10	0.70	0.52	0.08
Part-script total	27	0.23	0.78	0.66	0.07

Under a consensual definition of 'true' score, the lowest correlation achieved by a single participant was 0.09 on item 2a (Table 5). The highest was 0.93 on the part-script total. The part-script total exhibited the highest mean correlation and item 1a the lowest, 0.81 and 0.67 respectively. It is notable that all correlations are higher when the mean mark awarded by all participants is regarded as the 'true' score, rather than the mark awarded by the Principal Examiner. Of course, this is at least partly explained by the participants' marking feeding into the consensual 'true' mark. The standard deviation for each item and the part-script total were virtually identical.

Table 6 contains descriptive statistics for the absolute difference in marks awarded to work by the Principal Examiner and the participants as a percentage of the maximum absolute mark difference possible. Since after training participants marked a further 99 scripts, the maximum absolute mark difference possible for item 1a was 297 (3X99), for item 1b it was 396 (4X99),

and so on. Again, there was variation in the extent to which the participants' marking was similar to that of the Principal Examiner. On average, items 2b and 1c were the most discrepant and the part-script total the least. Item 1a exhibited the greatest variation in deviation from the mean absolute mark difference from the Principal Examiner and the part-script total the smallest.

Table 5 Descriptive statistics for the correlation between the mean marks awarded by all participants and those awarded by individual participants

Item	Maximum mark possible	Minimum	Maximum	Mean	Std. Deviation
1a	3	0.16	0.86	0.67	0.10
1b	4	0.34	0.87	0.72	0.07
1c	6	0.21	0.87	0.68	0.09
2a	8	0.09	0.89	0.78	0.08
2b	6	0.26	0.89	0.74	0.08
Part-script total	27	0.31	0.93	0.81	0.07

Table 6 Descriptive statistics for the absolute difference in marks awarded to work by the Principal Examiner and the participants as a percentage of the maximum absolute mark difference possible

Item	Maximum mark possible	Minimum	Maximum	Mean	Std. Deviation
1a	3	9.09	27.27	14.66	3.07
1b	4	6.31	22.47	11.50	2.22
1c	6	11.45	25.25	15.45	2.01
2a	8	8.96	20.33	13.69	1.90
2b	6	11.11	22.90	15.55	2.19
Part-script total	27	7.18	19.38	10.12	1.72

When the mean mark awarded to work by all participants was defined as the 'true' score, rather than that assigned by the Principal Examiner, item 1a exhibited the highest mean absolute mark difference as a percentage of the maximum absolute mark difference possible, 14 *percent*. Again the lowest mean absolute mark difference was achieved on the part-script total, on average, participants' marks disagreed with the mean mark awarded by all participants by 8 *percent* of the maximum absolute mark difference possible (Table 7). The spread of the absolute mark differences was greatest on item 1a and smallest on item 1b. With the exception of item 1b, the mean figures indicate lower discrepancies than when the Principal Examiner's mark is regarded as the 'true' score.

Descriptive analyses have shown that the correlations are maximised and absolute mark differences (as a percentage of the maximum absolute mark difference possible) minimised at the level of part-script total, a consequence of marking inconsistencies at item level cancelling each other out. On average, a consensus view of 'true' score results in higher correlations and smaller discrepancies in terms of absolute marks than when the Principal Examiner's judgement

is regarded as the 'gold standard'. It is arguable which of these is the best estimate of the candidates' 'true' score but it is clear that the Principal's marks varied from the average of the participants' marks. The closer the hierarchical and consensual operationalisations of 'true' mark the more successfully the Principal has conveyed their conceptualisation of the mark scheme.

Table 7 Descriptive statistics for the absolute difference in mean marks awarded to work by all the participants and individual participants as a percentage of the maximum absolute mark difference possible

Item	Maximum mark possible	Minimum	Maximum	Mean	Std. Deviation
1a	3	10.09	32.17	14.22	2.55
1b	4	7.84	22.00	11.75	1.97
1c	6	6.89	25.14	12.12	2.39
2a	8	6.02	20.42	10.65	2.29
2b	6	7.85	20.48	12.01	2.44
Part-script total	27	4.59	16.59	8.01	2.05

The effect of marker background and training on the reliability of marking

Using the four operationalisations of marking reliability, the effects of marker background and training on reliability were investigated using analyses of variance.

Item 1a

Principal Examiner's mark as the 'true' score - correlation between the Principal Examiner's and participants' marking of candidates' work

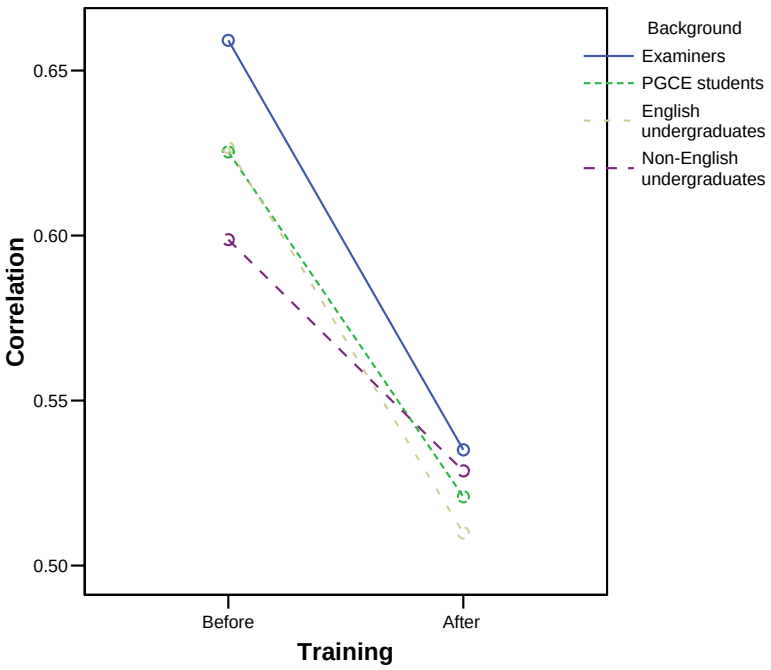
There was a significant effect of marker background on the correlation between the participants' marking of this item and that of the Principal Examiner ($F(3, 353) = 3.872$, $MS = 0.110$, $p=0.010$). Tukey *post-hoc* contrasts showed that examiners marked this item significantly more reliably than both groups of undergraduates. There was no significant difference in the reliability of the marking of PGCE students and examiners. Surprisingly training significantly reduced the correlation between the participants' marking of the item and that of the Principal ($F(1, 353) = 289.091$, $MS = 4.755$, $p<0.001$). This detrimental effect was not equal across the groups ($F(3, 353) = 3.435$, $MS = 0.056$, $p=0.017$). Training affected the undergraduates least. This group were the least reliable before training (see Figure 3).

Consensual view of 'true' score - correlation between the mean of marks awarded by all participants and the marks awarded by individual participants to candidates' work

Using a consensual rather than hierarchical definition of 'true' mark had little effect on the findings. There was a similar significant effect of marker background on the correlation between the participants' marking of the scripts and the mean mark of all participants ($F(3, 353) = 3.215$, $MS = 0.146$, $p=0.023$). Tukey *post-hoc* contrasts showed that examiners marked this item significantly more reliably than undergraduates did. Training significantly reduced the correlation

(F(1, 353) = 59.687, MS = 1.237, p<0.001) but in this case the detrimental effect was equal across the groups (F(3, 353) = 1.980, MS = 0.041, p=0.117) (see Figure 4).

Figure 3 The effect of marker background and training on the correlation between the Principal Examiner's and participants' marking of candidates' responses to item 1a



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

Depending on how reliability was operationalised, training and background seemed to have different effects. There was no significant main effect of marker background on the absolute difference in marks awarded to responses to this question by the Principal Examiner and the participants (F(3, 353) = 1.203, MS = 164.393, p=0.312). Nor was there a significant main effect of training on reliability (F(1, 353) = 1.056, MS = 83.539, p=0.307). There was a marginally significant interaction effect (F(3, 353) = 2.543, MS = 201.244, p=0.060) 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 English undergraduates and undergraduates (who marked relatively unreliably before training) (see Figure 5). The different findings when absolute mark difference was used to measure reliability rather than a correlation definition were partly caused by less statistical power for the analyses based on absolute mark difference. To be included in the analysis, participants needed to mark all 200 scripts. Only 14 of the PGCE students achieved this, for example[2].

[2] An analysis of the pattern of 'missing data' is underway.

Figure 4 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 1a

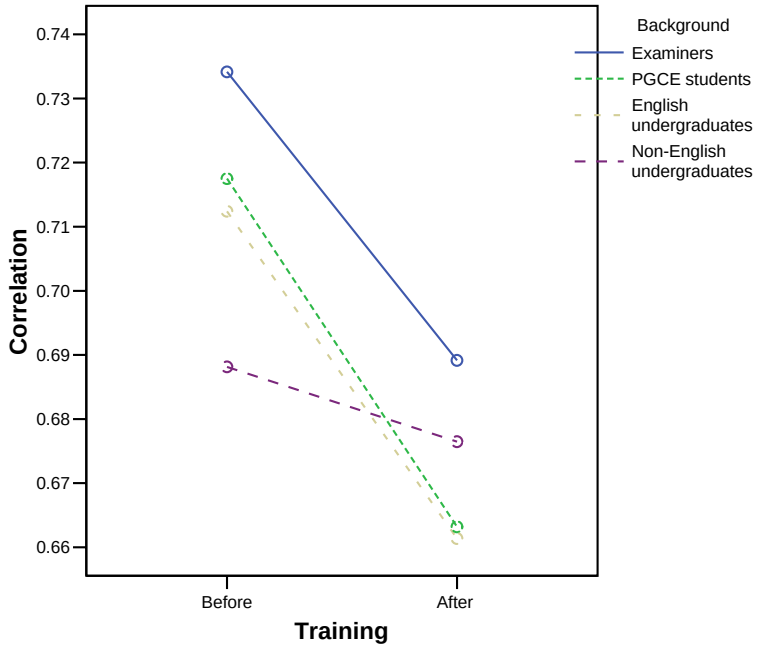
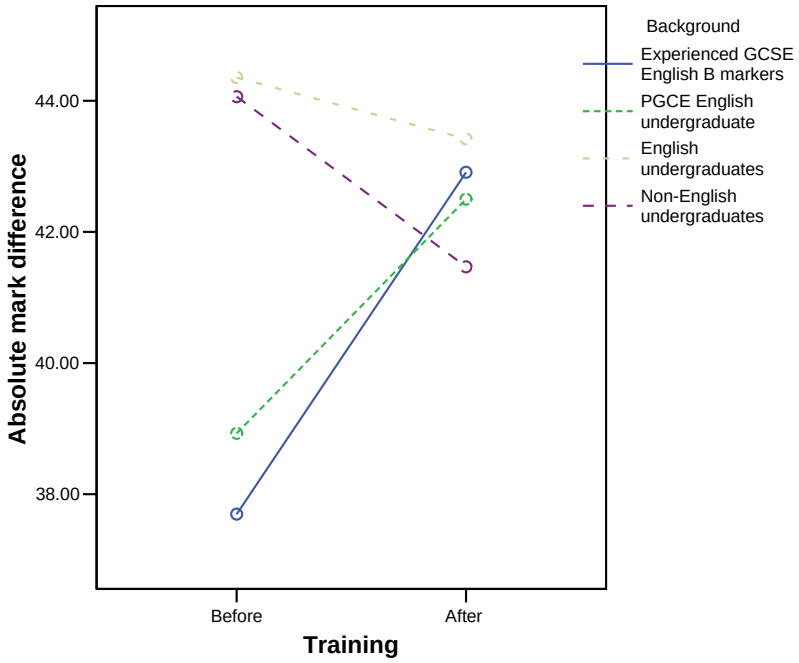


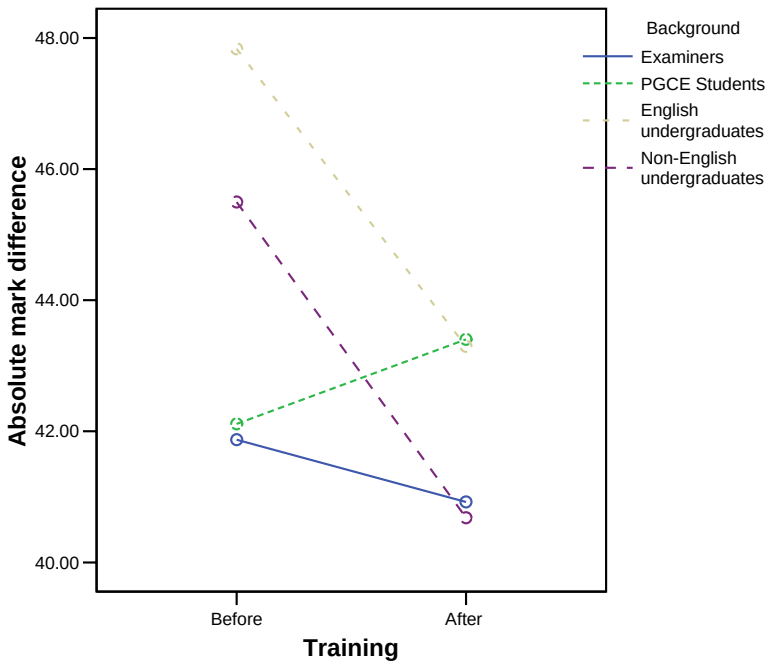
Figure 5 The effect of marker background and training on the absolute difference in marks awarded by the Principal Examiner and the participants to item 1a



Consensual view of 'true' score - absolute difference in the mean mark awarded to work by all the participants and that awarded by individual participants

Marker background had no effect on the absolute difference in marks awarded to responses to this question by participants and the consensual 'true' mark (F(3, 353) = 2.053, MS =188.989, p=0.111). There was, however, a significant positive impact of training on absolute mark difference which was not the case when 'true' mark was defined hierarchically (F(1, 353) = 5.567, MS = 245.598, p=0.020). Although Figure 6 shows that the PGCE students were the only group whose marking deteriorated following training, the effect of training was not *significantly* different for participants with different backgrounds (F(3, 353) = 2.155, MS = 95.061, p=0.098). The positive effect of training on absolute mark difference is at odds with the negative impact it had on the correlation between participants' marking and both definitions of 'true' mark.

Figure 6 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



In summary, when reliability was defined as the correlation between participants' marking and the 'true' score (whether defined hierarchically or consensually), examiners' marking was more reliable than that of undergraduates and English undergraduates. There was no difference between the marking reliability of the examiners and PGCE students.

However, when reliability was measured by the absolute difference in marks awarded by the participants and the 'true' score, background did not affect reliability, whichever definition of 'true' score was used. This may have been due to a lack of statistical power caused by few participants completing the marking of all the scripts. This had the effect of fewer cases being included in the analyses based on absolute mark difference than in those based on correlation.

It is also problematic to use correlations with such a restricted range of marks (0-3). This reduces the statistical power of these analyses leading to an effect being missed (Type II error). It is not clear, however, how restricted range might lead to a spurious effect with this data (Type 1 error).

Training reduced the correlation between participants' marks and the 'true' marks (whether defined hierarchically or consensually) but it also reduced the absolute mark difference from the consensual measure of 'true' mark (it had no effect on the absolute mark difference from the hierarchical 'true' mark). One would expect a reduction in absolute mark difference to be associated with an increase in the correlation. The effect of training on the range of marks awarded explains this seemingly contradictory finding. Participants were more likely to award extreme marks such as 0 or 3 before training. Training had the effect of reducing the spread of marks awarded, although not on the examiners marking. This can be seen from examination of the standard deviation of marks before and after training (see Table 8). This effect 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.

Table 8 The standard deviation of marks awarded before and after training

Background	Before training	After training
Examiners	25.80	25.36
PGCE students	29.05	19.84
English undergraduates	33.60	19.08
Undergraduates	25.68	20.92

Item 1b

Principal Examiner's mark as the 'true' score - correlation between the Principal Examiner's and participants' marking of candidates' work

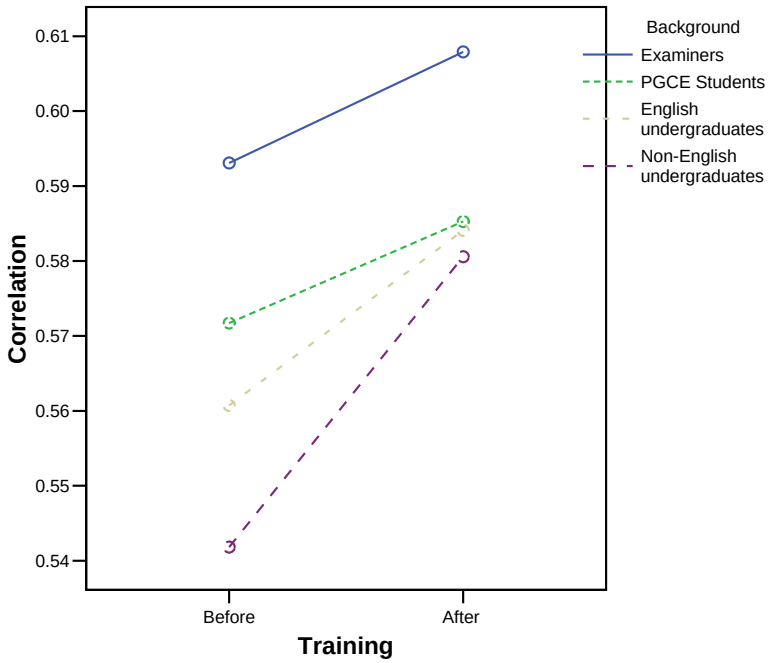
There was a significant effect of marker background on the correlation between the participants' marking of the scripts and that of the Principal Examiner ($F(3, 353) = 6.571$, $MS = 0.112$, $p<0.001$). Tukey *post-hoc* contrasts showed that examiners marked this item significantly more reliably than both groups of undergraduates. There was no significant difference in the reliability of the marking of PGCE students and examiners. This is the same pattern of findings as for item 1a. For this item, however, training significantly improved the correlation between the participants' marking of the scripts and that of the Principal Examiner ($F(1, 353) = 20.639$, $MS = 0.201$, $p<0.001$). This effect was not significantly different across the groups ($F(3, 353) = 0.874$, $MS = 0.009$, $p=0.455$) (see Figure 7).

Consensual view of 'true' score - correlation between the mean of marks awarded by all participants and the marks awarded by individual participants to candidates' work

There was a significant effect of marker background on the correlation between the participants' marking of the scripts and the mean of all participants' marking ($F(3, 353) = 7.590$, $MS = 0.240$, $p<0.001$). Tukey *post-hoc* contrasts showed that examiners marked this item significantly more reliably than all the other groups of participants did. Training significantly reduced the correlation ($F(1, 353) = 108.786$, $MS = 1.379$, $p<0.001$). This detrimental effect was not equal across the groups ($F(3, 353) = 5.703$, $MS = 0.072$, $p=0.001$) (see Figure 8). It reduced the quality of

marking of the PGCE students most, who were marking almost as reliably as the examiners before training. After training, however, they became the least reliable group.

Figure 7 The effect of marker background and training on the correlation between the Principal Examiner's and participants' rank marking of candidates' responses to item 1b



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

There was a significant effect of marker background on the absolute difference in marks awarded to responses to this question by the Principal Examiner and the participants ($F(3, 353) = 3.960$, $MS = 365.797$, $p=0.010$). *Post hoc* Tukey contrasts showed that the PGCE students were significantly more reliable than the English undergraduates were. There was a significant positive impact of training on reliability ($F(1, 353) = 208.123$, $MS = 11154.533$, $p<0.001$). This effect was equal across the groups of participants ($F(3, 353) = 0.802$, $MS = 42.973$, $p=0.496$) (see Figure 9).

Figure 8 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 1b

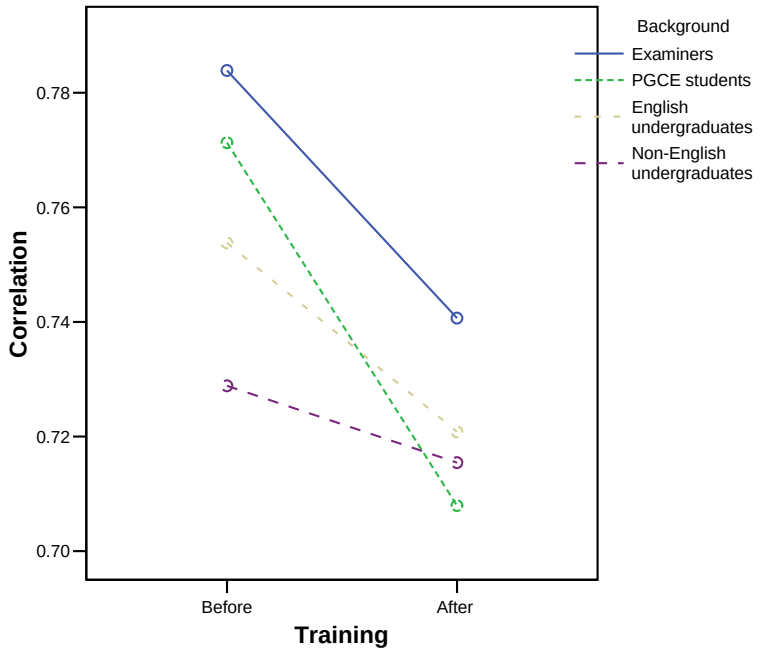
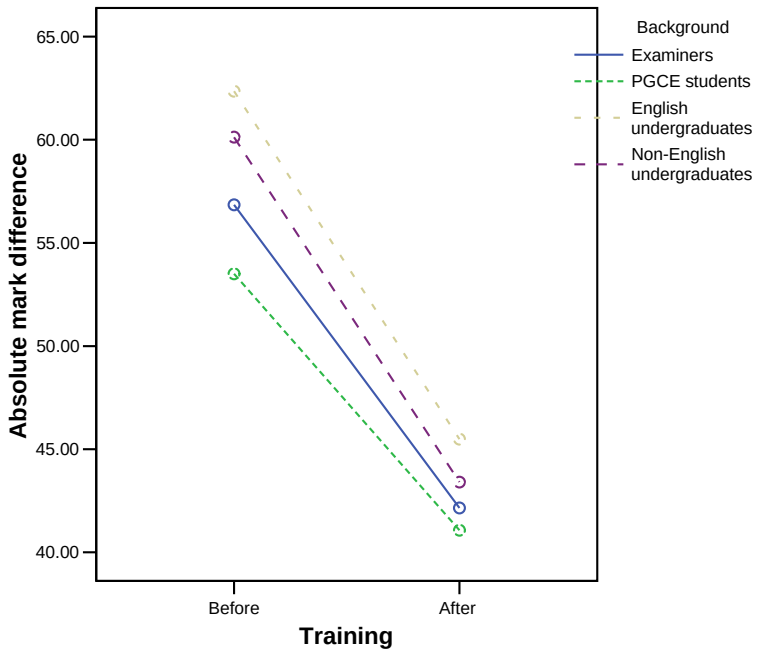


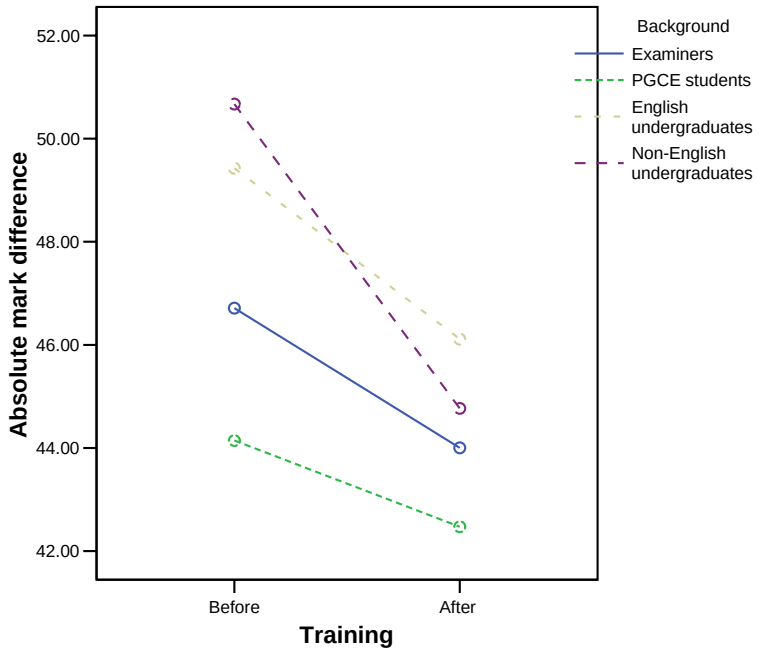
Figure 9 The effect of marker background and training on the absolute difference in marks awarded by the Principal Examiner and the participants to item 1b



Consensual view of 'true' score - absolute difference in marks awarded to work by the Principal Examiner and the participants

There was no significant effect of marker background on the absolute difference in marks awarded to responses to this question by the participants and the mean mark of all participants ($F(3, 353) = 2.158$, $MS = 188.574$, $p=0.097$). There was a significant positive impact of training on reliability ($F(1, 353) = 14.495$, $MS = 559.340$, $p<0.001$). The effect of training was not significantly different for participants with different backgrounds ($F(3, 353) = 1.079$, $MS = 41.623$, $p=0.361$) (see Figure 10).

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



As for item 1a, the effects of training and background depended on the definition of marking reliability. Training reduced the absolute mark difference from the 'true' mark (whether defined hierarchically or consensually) and increased the correlation between participants' marks and the 'true' marks when defined hierarchically. However, training decreased the correlation between participants' marks and the 'true' marks when defined consensually but this effect was only statistically significant for the PGCE students.

Using the latter consensual definition of reliability, examiners marked more reliably than all the other groups of participants, including the PGCE students. However, this was caused by the adverse impact of the training on the PGCE students' marking.

Examiners were no more or less reliable in terms of the absolute difference in marks from the hierarchical 'true' mark than the other groups but their marking correlated with the hierarchical 'true' mark better than that of both groups of undergraduates, although no better than that of the

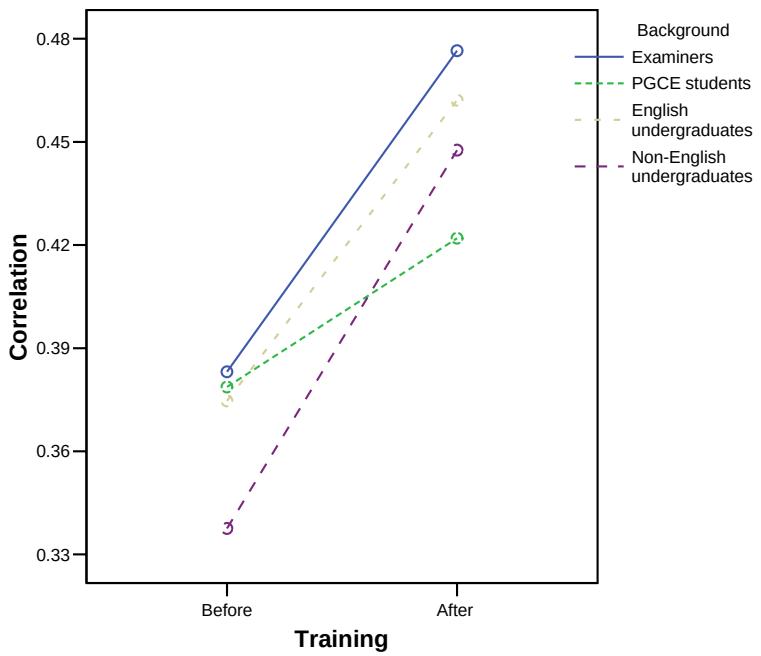
PGCE students. PGCE students' marking also had lower absolute difference in marks than the English undergraduates' marking. However, there was no significant effect of background on the absolute difference in marks from the consensual 'true' mark.

Item 1c

Principal Examiner's mark as the 'true' score - correlation between the Principal Examiner's and participants' marking of candidates' work

For this item too training significantly improved the correlation between the participants' marking of the scripts and that of the Principal Examiner ($F(1, 353) = 213.489$, $MS = 1.827$, $p<0.001$). This effect, however, was not equal across the groups ($F(3, 353) = 5.801$, $MS = 0.050$, $p<0.001$). The PGCE students benefited from the training significantly less than other participants did. There was a significant effect of marker background on the correlation between the participants' marking of the scripts and that of the Principal ($F(3, 353) = 6.799$, $MS = 0.072$, $p<0.001$). Tukey *post-hoc* contrasts showed that the examiners marked significantly more reliably than the undergraduates and PGCE students. The difference between the marking reliability of the examiners and PGCE students was due to the lack of improvement in their marking following training (see Figure 11).

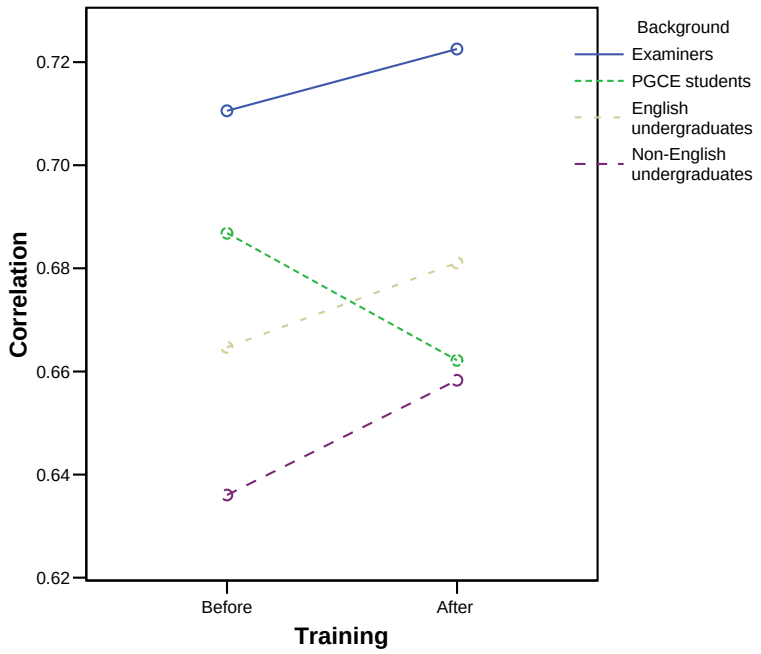
Figure 11 The effect of marker background and training on the correlation between the Principal Examiner's and participants' marking of candidates' responses to item 1c



Consensual view of 'true' score - correlation between the mean of marks awarded by all participants and the marks awarded by individual participants to candidates' work

There was a significant effect of marker background on the correlation between the participants' marking of the scripts and the mean mark of all participants (F(3, 353) = 14.201, MS = 0.521, p<0.001). Tukey *post-hoc* contrasts showed that examiners marked this item significantly more reliably than all the other groups of participants did. Overall training significantly improved the correlation (F(1, 353) = 1.576, MS = 0.025, p=0.210) but it had a negative effect on the quality of the marking done by the PGCE students (F(3, 353) = 4.079, MS = 0.064, p=0.007) (see Figure 12).

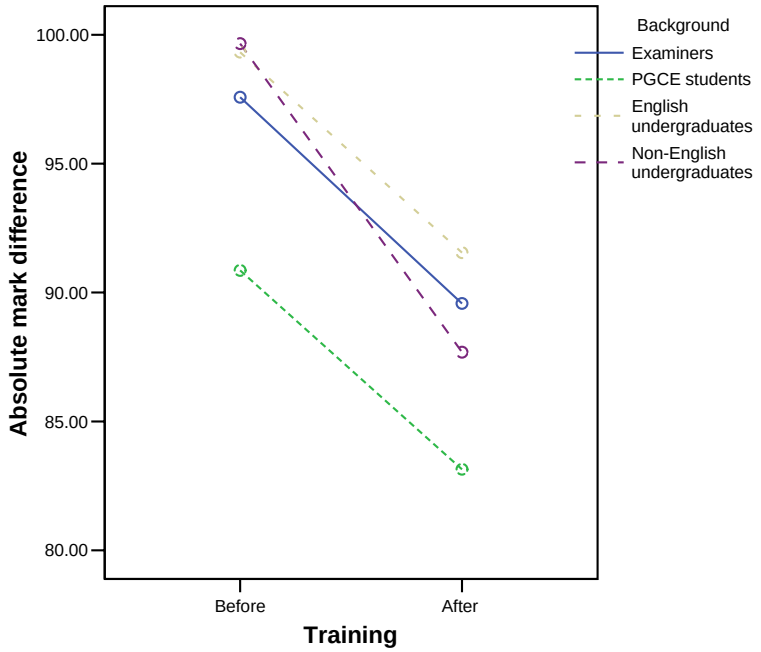
Figure 12 The effect of marker background and training on the correlation between the mean of marks awarded by all participants and the marks awarded by individual participants to candidates' responses to item 1c



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

There was no significant effect of marker background on the absolute difference in marks awarded to responses to this question by the Principal and the participants (F(3, 353) = 2.188, MS =469.303, p=0.094). There was a significant positive impact of training on reliability (F(1, 353) = 39.660, MS = 3804.217, p<0.001). There was no significant interaction between the impact of the training and the background of the participants (F(3, 353) = 0.668, MS = 64.071, p=0.574) (see Figure 13).

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 1c



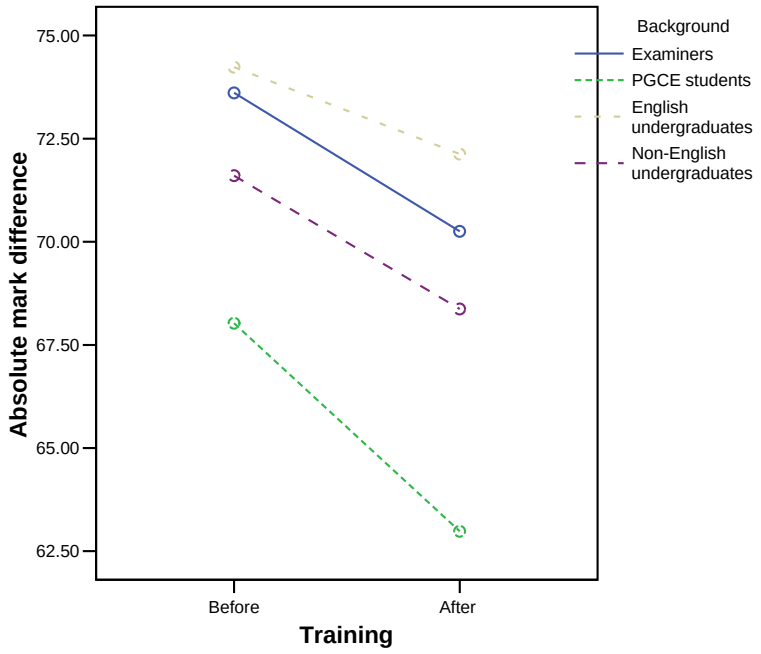
Consensual view of 'true' score - absolute difference in marks awarded to work by the Principal Examiner and the participants

There was no significant effect of marker background on the absolute difference in marks awarded to responses to this question by the participants and the mean mark of all participants ($F(3, 353) = 1.194$, $MS = 420.120$, $p=0.316$). There was a significant positive impact of training on reliability ($F(1, 353) = 5.381$, $MS = 571.375$, $p=0.022$) which was not significantly different for participants with different backgrounds ($F(3, 353) = 0.134$, $MS = 14.245$, $p=0.940$) (see Figure 14).

In summary, training reduced the absolute difference from the 'true' mark (whether defined hierarchically or consensually). Overall, training also had a positive impact on the correlation between participants' marks and the 'true' mark (again whether defined hierarchically or consensually) but not for the PGCE students. For these participants, training had no significant impact on the correlation with the hierarchical 'true' mark and actually reduced the correlation with the consensual 'true' mark.

There was no effect of marker background on the absolute difference in marks from the 'true' mark (whether defined hierarchically or consensually). There was however an effect of background on the correlation definition of reliability. Examiners' marking correlated better with the hierarchical definition of 'true' mark than that of undergraduates or PGCE students, and better with the consensual definition of 'true' mark than the marking of all the other groups. This gap in marking reliability between the PGCE students and the examiners seemed to be due to the detrimental effect of training on PGCE students' marking.

Figure 14 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

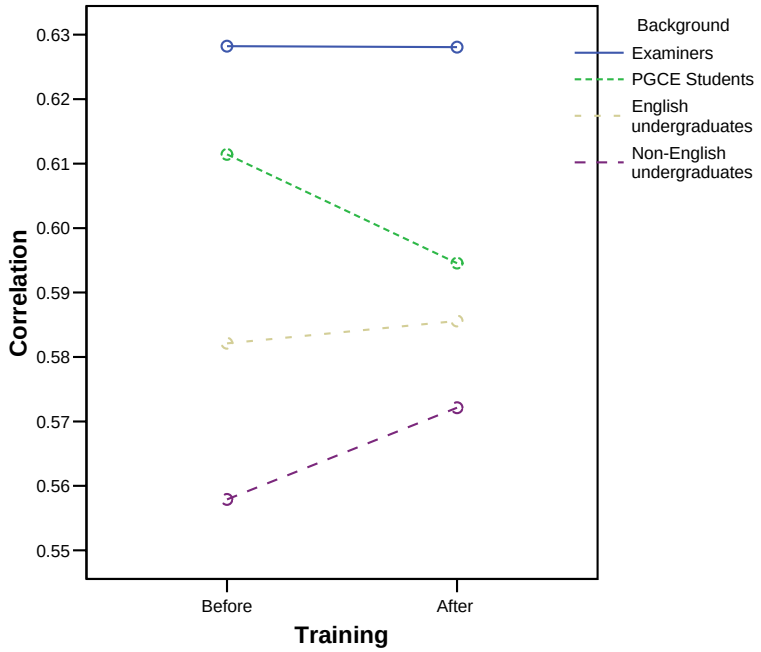


Item 2a

Principal Examiner's mark as the 'true' score - correlation between the Principal Examiner's and participants' marking of candidates' work

For this item training had no significant impact on the correlation between the participants' marking of the scripts and that of the Principal ($F(1, 353) = 0.025$, MS <0.001, $p=0.873$). This was true for all the groups ($F(3, 353) = 1.291$, MS $= 0.012$, $p=0.277$). There was however a significant effect of marker background on reliability ($F(3, 353) = 15.837$, MS $= 0.287$, $p<0.001$). Tukey *post-hoc* contrasts showed that the examiners marked significantly more reliably than all the other groups and that PGCE students marked more reliably than the undergraduates (see Figure 15).

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



Consensual view of 'true' score - correlation between the mean of marks awarded by all participants and the marks awarded by individual participants to candidates' work

There was a significant effect of marker background on the correlation between the participants' marking of the scripts and the consensual 'true' mark ($F(3, 353) = 22.073$, MS = 0.980, $p<0.001$). Tukey *post-hoc* contrasts showed that examiners marked this item significantly more reliably than all the other groups of participants did and that the PGCE students marked more reliably then the undergraduates. Overall training significantly improved the correlation ($F(1, 353) = 49.318$, MS = 0.757, $p<0.001$) but it had very little impact on the quality of the marking done by the PGCE students ($F(3, 353) = 5.227$, MS = 0.080, $p=0.002$) (see Figure 16).

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

There was a significant effect of marker background on the absolute difference in marks awarded to responses to this question by the Principal Examiner and the participants ($F(3, 353) = 3.806$, MS =4929.220, $p=0.012$). Tukey *post hoc* tests were not significant but the examiners and PGCE students had lower mark differences than both groups of undergraduates. There was a significant positive impact of training on reliability ($F(1, 353) = 7.362$, MS = 1293.631, $p=0.008$) which did not interact with the background of the participants ($F(3, 353) = 0.047$, MS = 8.201, $p=0.987$) (see Figure 17).

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

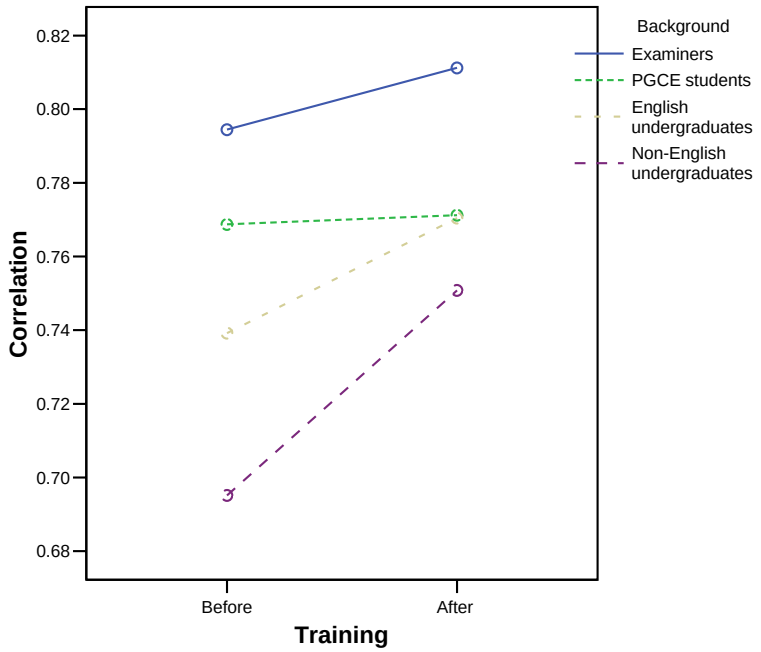
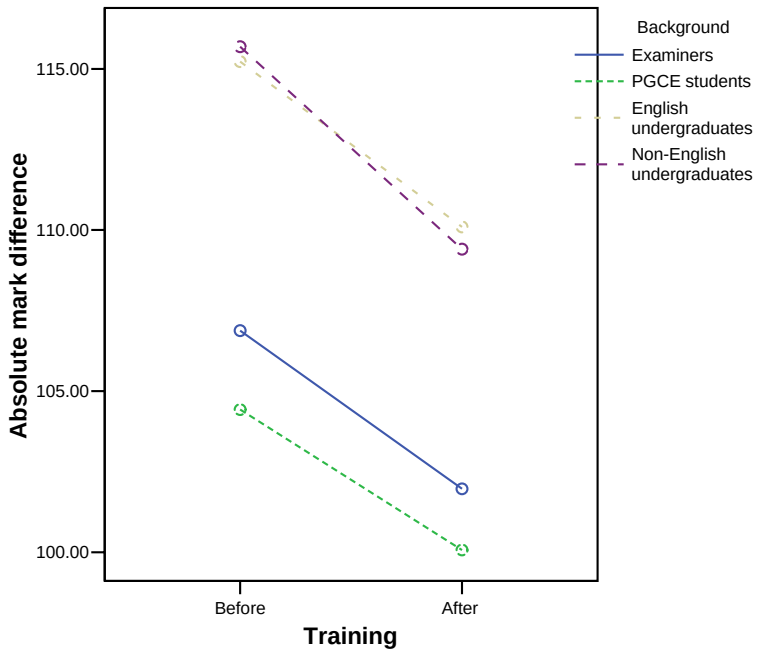


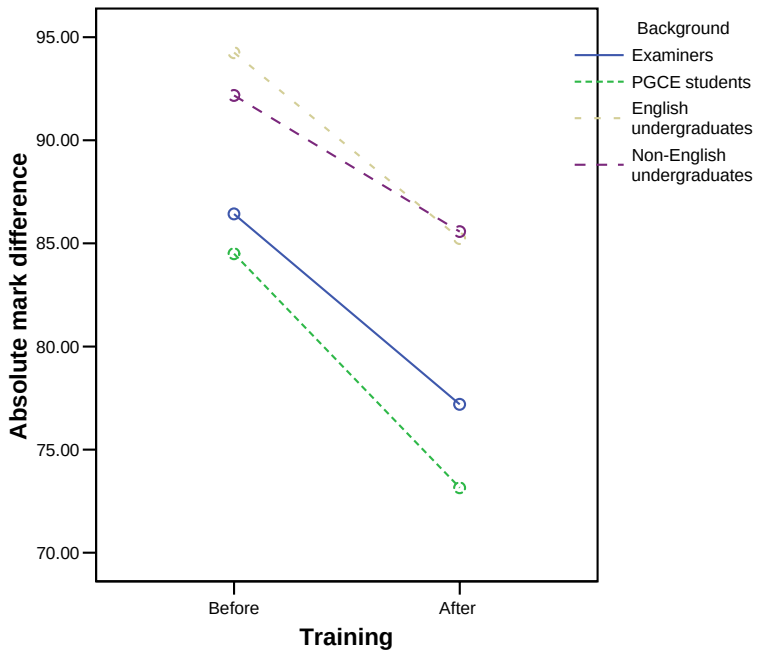
Figure 17 The effect of marker background and training on the absolute difference in marks awarded by the Principal Examiner and the participants to item 2a



Consensual view of 'true' score - absolute difference in marks awarded to work by the Principal Examiner and the participants

There was a marginally significant effect of marker background on the absolute difference in marks awarded to responses to this question by all the participants and that awarded by individual participants (F(3, 353) = 2.720, MS =1326.045, p=0.048). Tukey *post hoc* contrasts were non-significant but Figure 18 shows 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 (F(1, 353) = 17.584, MS = 3963.452, p<0.001). The effect of training was not significantly different for participants with different backgrounds (F(3, 353) = 0.186, MS = 41.897, p=0.906).

Figure 18 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



In summary, training reduced the absolute difference from the 'true' mark (whether defined hierarchically or consensually). Overall, training also had a positive impact on the correlation between participants' marks and the consensual 'true' mark although it had no impact on the quality of marking of the PGCE students. It had no significant effect on the correlation between participants' marks and the hierarchical 'true' mark.

Examiners' marking correlated better with the 'true' mark (hierarchical or consensual) than that of all other groups, and PGCE students' marking correlated better than that of the undergraduates. Again, the gap in marking reliability between the PGCE students and the examiners may be due to training having little positive impact on PGCE students' marking. There were also significant main effects of background on absolute difference from both definitions of 'true' mark. The *post hoc* analyses of these main effects were non-significant but

the marking of the examiners and PGCE students tended to be more reliable than that of both groups of undergraduates.

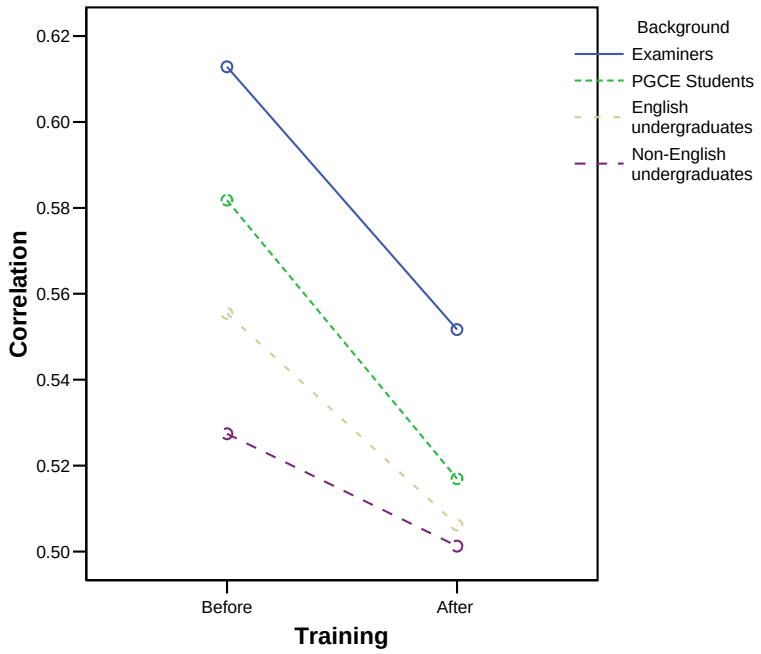
Item 2b

Principal Examiner's mark as the 'true' score - correlation between the Principal Examiner's and participants' marking of candidates' work

There was a significant effect of marker background on the correlation between the participants' marking of the scripts and that of the Principal (F(3, 353) = 18.413, MS = 0.310, p<0.001). Tukey *post-hoc* contrasts showed that examiners marked this item significantly more reliably than the other groups of markers and that PGCE students marked more reliably than undergraduates did.

Training significantly reduced the correlation between the participants' marking of the scripts and that of the Principal (F(1, 353) = 103.551, MS = 1.015, p<0.001). This detrimental effect was not equal across the groups (F(3, 353) = 3.461, MS = 0.034, p=0.017). Training affected the undergraduates least. This group were the least reliable before training (see Figure 19).

Figure 19 The effect of marker background and training on the correlation between the Principal Examiner's and participants' marking of candidates' responses to item 2b

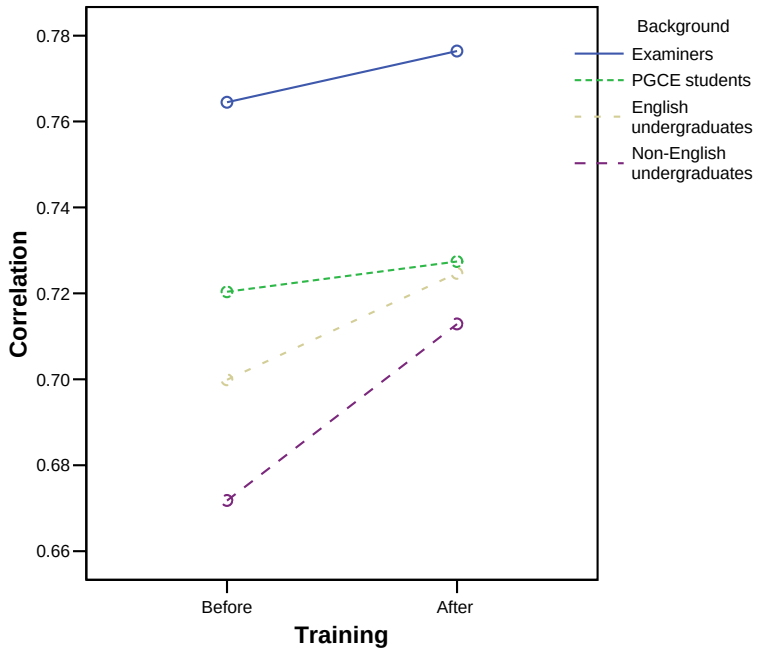


Consensual view of 'true' score - correlation between the mean of marks awarded by all participants and the marks awarded by individual participants to candidates' work

There was a significant effect of marker background on the correlation between the participants' marking of the scripts and the mean of all participants' marking (F(3, 353) = 25.277, MS = 0.903, p<0.001). Tukey *post-hoc* contrasts showed that examiners marked this item significantly

more reliably than all the other groups of participants did and that the PGCE students marked more reliably then the undergraduates. Overall training significantly improved the correlation (F(1, 353) = 25.675, MS = 0.389, p<0.001) and this effect wasn't significantly different across the groups (F(3, 353) = 2.256, MS = 0.034, p=0.082) (see Figure 20).

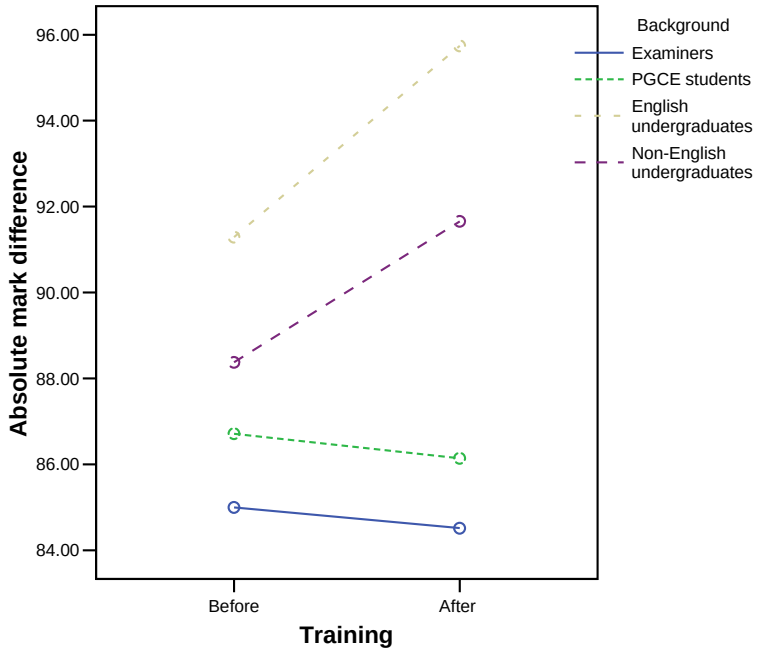
Figure 20 The effect of marker background and training on the correlation between the mean of marks awarded by all participants and the marks awarded by individual participants to candidates' responses to item 2b



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

There was a significant effect of marker background on the absolute difference in marks awarded to responses to this question by the Principal and the participants (F(3, 353) = 2.989, MS =876.947, p=0.034). Tukey *post hoc* tests showed that the examiners had lower mark differences than the English undergraduates. There was no significant impact of training on reliability (F(1, 353) = 1.054, MS = 133.347, p=0.307). This was the case no matter what the background of the participants (F(3, 353) = 0.688, MS = 87.089, p=0.561) (see Figure 21).

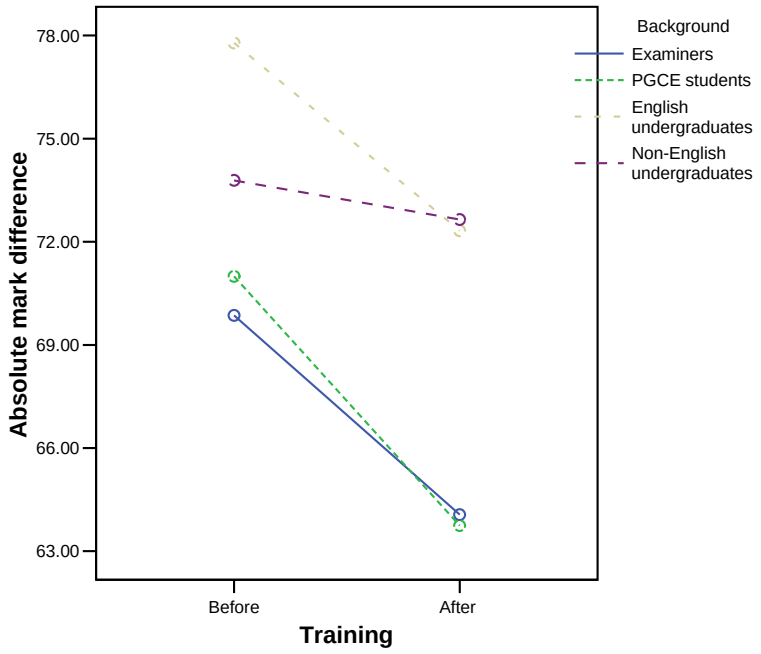
Figure 21 The effect of marker background and training on the absolute difference in marks awarded by the Principal Examiner and the participants to item 2b



Consensual view of 'true' score - absolute difference in marks awarded to work by the Principal Examiner and the participants

There was no significant effect of marker background on the absolute difference in marks awarded to responses to this question by all the participants and that awarded by individual participants (F(3, 353) = 2.421, MS =901.919, p=0.070) (see Figure 22). There was a significant positive impact of training on reliability (F(1, 353) = 8.239, MS = 1154.863, p=0.005). The effect of training was not significantly different for participants with different backgrounds (F(3, 353) = 0.625, MS = 91.388, p=0.583).

Figure 22 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



In summary, training had a positive impact on absolute mark difference and the correlation when the consensual definition of 'true' mark was employed. However, it had no impact on the absolute difference in marks from the hierarchical definition of 'true' mark and had a negative effect on the size of the correlation with the hierarchical 'true' mark. In the latter case, training had the least detrimental effect on the undergraduates' marking, whose marking was least reliable before training.

As with item 1a, participants were more likely to award extreme marks before than after training. This can be seen from examination of the standard deviation of marks before and after training (Table 9).

Table 9 The standard deviation of marks awarded before and after training

Background	Before training	After training
Examiners	38.20	34.68
PGCE students	43.30	35.01
English undergraduates	48.68	43.84
Undergraduates	43.38	40.89

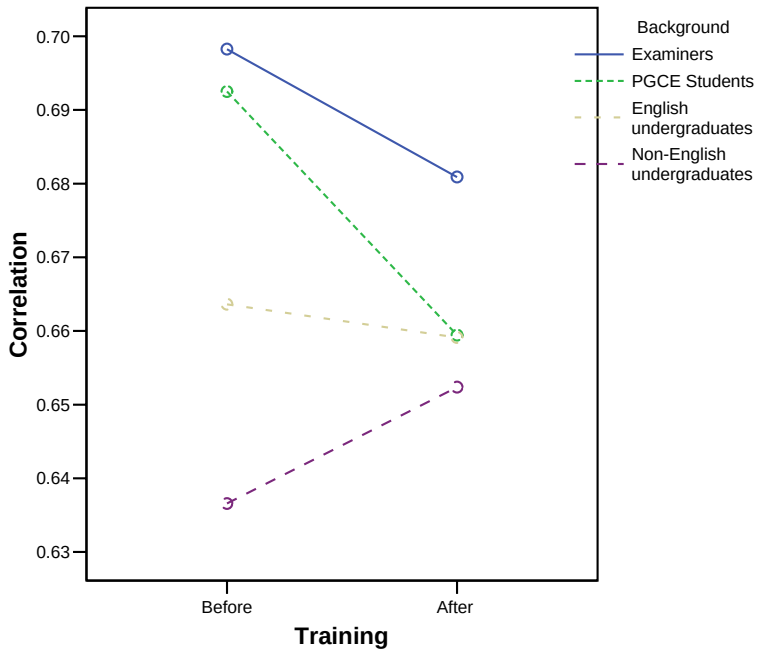
Examiners' marking correlated better with the 'true' mark (hierarchical or consensual) than that of all other groups, and PGCE students' marking correlated better than that of the undergraduates'. Examiners' marking was also closer than that of the English undergraduates in terms of absolute mark difference from the hierarchical 'true' mark. However, there was no significant effect of background on absolute mark difference from the consensual 'true' mark.

Part-script total

Principal Examiner's mark as the 'true' score - correlation between the Principal Examiner's and participants' marking of candidates' work

There was a significant effect of marker background on the correlation between the participants' marking of the scripts and that of the Principal (F(3, 353) = 11.146, MS = 0.196, p<0.001). Tukey *post-hoc* contrasts showed that examiners marked significantly more reliably than the English undergraduates and undergraduates, and that PGCE students marked more reliably than undergraduates did. Overall training significantly reduced the correlation between the participants' marking of the scripts and that of the Principal (F(1, 353) = 7.364, MS = 0.074, p=0.007). Training had a differential effect across the groups (F(3, 353) = 5.743, MS = 0.058, p<0.001). Simple effect analysis showed that training had no significant effect on the undergraduates or English undergraduates, while it reduced the marking reliability of the examiners and PGCE students who had relatively high correlation coefficients prior to training (see Figure 23).

Figure 23 The effect of marker background and training on the correlation between the Principal Examiner's and participants' marking of the part-scripts

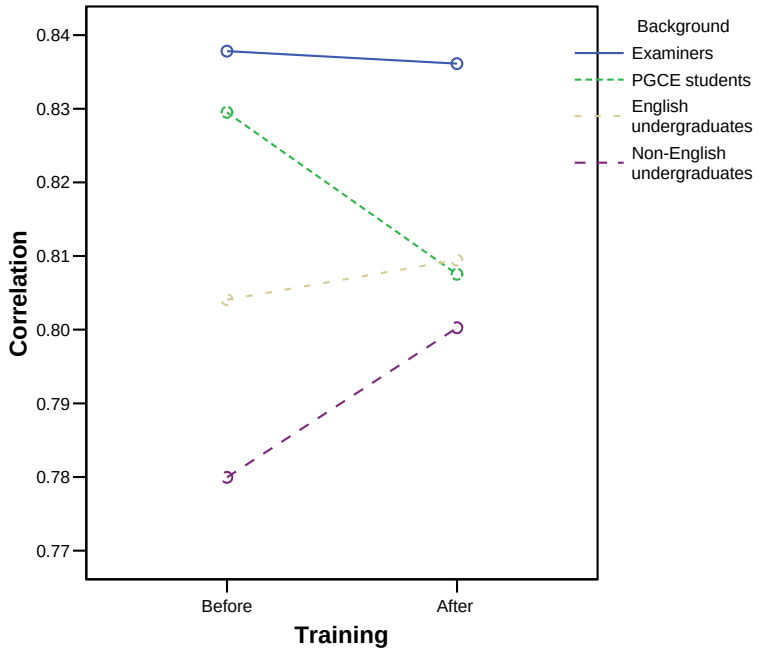


Consensual view of 'true' score - correlation between the mean of marks awarded by all participants and the marks awarded by individual participants to candidates' work

There was a significant effect of marker background on the correlation between the participants' marking of the scripts and the mean mark of all participants (F(3, 353) = 11.754, MS = 0.541, p<0.001). Tukey *post-hoc* contrasts showed that examiners marked this item significantly more reliably than both groups of undergraduates did. Further, the PGCE students marked more

reliably than the undergraduates did. While there was no significant main effect of training on the correlation (F(1, 353) = .044, MS = 0.001, p=0.835), this varied by group (F(3, 353) = 5.340, MS = 0.092, p=0.001) (Figure 24). Training improved the marking of the English undergraduates and undergraduates, who tended to mark relatively unreliably before training. Training had little impact on the reliability of the marking of examiners but clearly reduced the quality of marking done by PGCE students.

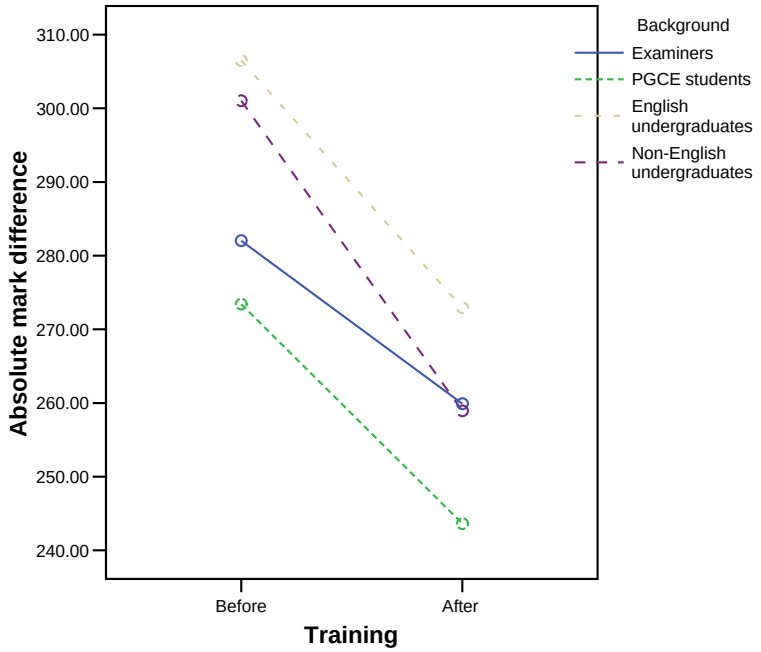
Figure 24 The effect of marker background and training on the correlation between the mean of marks awarded by all participants and the marks awarded by individual participants to candidates' responses to total part-script



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

There was no significant effect of marker background on the absolute difference in marks awarded to part-scripts by the Principal and the participants (F(3, 353) = 2.245, MS =7450.497, p=0.087). There was a significant positive impact of training on reliability (F(1, 353) = 29.386, MS = 48658.117, p<0.001). This was the case no matter what the background of the participants (F(3, 353) = 0.645, MS = 1068.843, p=0.588) (see Figure 25).

Figure 25 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

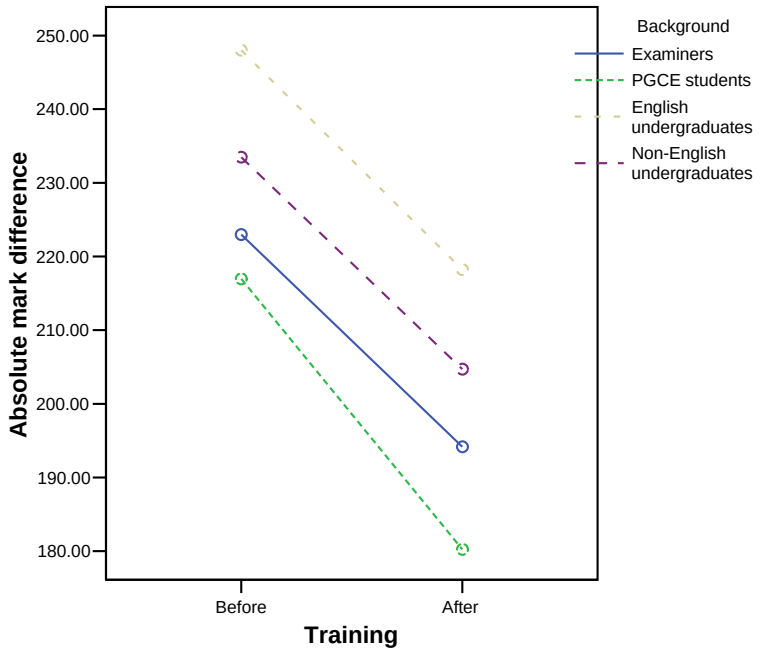


Consensual view of 'true' score - absolute difference in marks awarded by all participants and the marks awarded by individual participants to candidates' work

There was no significant effect of marker background on the absolute difference in marks awarded to the part-script by the participants and mean mark of participants ($F(3, 353) = 2.1666$, $MS = 10070.295$, $p=0.096$) (see Figure 26). There was a positive impact of training on reliability ($F(1, 353) = 22.584$, $MS = 46050.395$, $p<0.001$) which was not significantly different for participants with different backgrounds ($F(3, 353) = 0.059$, $MS = 120.662$, $p=0.981$).

In summary, there was no effect of background on the absolute difference in marks from either the hierarchical or the consensual measure of 'true' mark. However, background did have an effect on the correlation measure of reliability. For both measures of 'true' mark, examiners were more reliable than English undergraduates and undergraduates. PGCE students' marking was more reliable than that of undergraduates. Note there was no difference in the reliability of marking of PGCE students and examiners.

Figure 26 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



The effect of training on marking reliability depended on how reliability was operationalised. It reduced the absolute mark difference from both measures of 'true' mark. It had, however, a differential effect on the correlation reliability measure. For the examiners and PGCE students training reduced the correlation between participants' marks and the hierarchical 'true' marks, though their marking prior to training correlated relatively well with that of the Principal. However, training improved the correlation for both groups of undergraduates whose marking prior to training correlated relatively poorly with that of the Principal. Training had a similar effect on the correlations with the consensual 'true' mark. It improved the correlations of both groups of undergraduates (which were relatively low prior to training), had no effect on the correlations of the examiners' marking but reduced the correlations of the PGCE students' marking.

One might expect a reduction in absolute mark difference to be associated with an increase in the correlation. As discussed earlier in relation to the marking of items 1a and 2b, 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 participants' marking more cautious. This can be seen from examination of the standard deviation of marks before and after training (see Table 10). This effect 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.

Table 10 The standard deviation of marks awarded before and after training

Background	Before training	After training
Examiners	147.59	134.10
PGCE students	180.12	129.40
English undergraduates	193.01	154.13
Undergraduates	166.75	148.90

Discussion of the effect of marker background and training on the reliability of marking

The effect of background on the quality of participants' marking depended upon whether the correlation or absolute mark difference operationalisation of reliability was used, upon the definition of 'true' mark (consensual or hierarchical) and upon the item being marked. Nonetheless, the findings formed a pattern. When reliability was measured by the absolute difference in marks awarded by the participants and either definition of 'true' score, background rarely affected reliability. This may have been due to a lack of statistical power caused by few participants completing the marking of all the scripts. This meant that fewer cases were included in the analyses using absolute mark difference than in those using correlation. Where background had a significant impact (items 1b, 2a and 2b), the marking of examiners and PGCE students was significantly closer to the 'true' mark than that of English undergraduates or undergraduates.

When reliability was defined as the correlation between participants' marking and the 'true' score (whether defined hierarchically or consensually), examiners' marking of all items was more reliable than that of undergraduates and English undergraduates. This was also 'true' at the level of part-script. For some items (2a and 2b, and 1b only when the consensual 'true' mark was used) the examiners' marking also correlated better with the 'true' score than that of the PGCE students. These items require longer responses and so one might expect that more experience is required to mark them. There is no evidence to suggest that PGCE students could not be employed to mark short answer questions. Indeed, there was no difference in the reliability of marking of examiners and PGCE students at the level of part-script. The natural conclusion of this is that PGCE students could be employed to mark whole scripts but not items that require longer answers alone. The implications of such a conclusion are discussed later.

Moreover, the difference in the marking reliability of examiners and PGCE students on longer answer questions often seemed to be explained by the PGCE students' response to training. For these items, training had either no effect or a detrimental effect on their marking. The findings suggest that PGCE students need a different form of training from that currently used to standardise examiners' marking. Further work is needed to establish the form of that training. The qualitative evaluations of the training gathered from the PGCE students did not highlight any specific problems. Indeed, they gave mostly positive evaluations but said that they would have liked more. More qualitative work may be valuable which could lead to the testing of bespoke training programmes.

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. However, training had a positive effect on the strength of the correlation of marking with the consensual 'true' mark for this item. One would expect a reduction in absolute mark difference to be associated with an increase in the correlation, but participants were less likely to award extreme marks after training. This compression effect was replicated at part-script level. Again, the positive effect of

training in reducing the absolute mark difference from the 'true' mark was 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. Indeed training materials distributed to senior examiners refer to the desirability of encouraging a spread of marks, for example:

"It is important that, from the beginning of the marking process and at the pre-standardisation and full standardisation meetings in particular, all Team Leaders and Assistant Examiners are clear about the desirability of using the full range of marks" (AQA, 2003 p.8).

Predictors of the reliability of marking: the role of age, gender, background, attitude and personality

The independent predictors of marking reliability prior to training and following training were investigated. As for the previous analyses, four operationalisations of marking reliability were used: correlation between participants' marking and the 'true' score (defined hierarchically and consensually) and absolute mark difference from the 'true' score (defined hierarchically and consensually).

Stepwise multiple regression analyses investigated the independent predictors of marking reliability prior to training and post training. Measures of reliability were regressed onto participants' age, gender, background (whether they were examiners, PGCE students, English undergraduates or other undergraduates), attitude to marking (their enjoyment of marking, the belief that only teachers should mark and their view of the role of judgement in marking) and personality (that is their scores on neuroticism, extraversion, openness to experience, agreeableness and conscientiousness scales).

The reliability of marking prior to training

Principal Examiner's mark as the 'true' score - correlation between the marks awarded by the Principal Examiner and marks awarded by the participants

Prior to training, marker background was a significant independent predictor of marking reliability (see Table 11). Examiners tended to mark items 1a, 1b, 2a and 2b more reliably than participants with other backgrounds did. PGCE students also tended to mark items 2a and 2b more reliably than undergraduates or English undergraduates. On the other hand, undergraduates tended to mark items 1a, 1b and 1c less reliably than participants with other backgrounds did and this difference remained at the level of part-script. English undergraduates also appeared to be less reliable than PGCE students or examiners at the level of part-script.

Measures of personality also independently predicted marking reliability, although not always in a straightforward manner. One can imagine that the extent to which participants were open to new experiences would predict marking reliability, after all two thirds of the participants had no experience of marking. However, it did so in a contradictory manner. Relatively open participants tended to mark item 1a less reliably but marked item 2a more reliably. Agreeable, co-operative participants tended to mark item 1a more reliably than other participants did. Conscientious participants tended to mark item 1c reliably. Personality did not independently predict marking reliability at part-script level.

The extent to which participants reported enjoying marking predicted marking reliability for items 1c and 2b, with those participants who enjoyed marking tending to mark these items more reliably. There was also an effect of gender on the extent to which item 1a was marked reliably: male participants tended to mark this item more reliably than female participants.

While these relationships are interesting, it is worth noting the relatively small amount of variance in marking reliability measured by the variables. The greatest amount of variation in reliability accounted for was 12 *percent* for item 2b; 11 *percent* was accounted for at part-script level. This is partly due to the inherently noisy nature of marking consistency, and given the failure of the majority of past research to find relationships between examiner background and traits and marking reliability, it should not be interpreted too pessimistically.

Table 11 Independent predictors of the correlation between the Principal Examiner's and participants' marking of item and part-script responses

Variable	Beta	t	p
Item 1a **R^2=0.114, F(5,313)=8.079, p<0.001**			
Examiners	**0.20**	**3.42**	**0.001**
Openness	**-0.17**	**-3.15**	**0.002**
Agreeableness	**0.13**	**2.38**	**0.018**
Gender	**0.12**	**2.29**	**0.023**
Undergraduates	**-0.12**	**-2.06**	**0.040**
Item 1b **R^2=0.059, F(2,316)=9.944, p<0.001**			
Examiners	**0.17**	**2.97**	**0.003**
Undergraduates	**-0.13**	**-2.21**	**0.028**
Item 1c R^2=0.081, F(3,315)=9.282, p<0.001			
Enjoyment of marking	**0.15**	**2.68**	0.008
Undergraduates	**-0.16**	**-2.95**	0.003
Conscientiousness	**0.11**	**2.00**	0.046
Item 2a R^2=0.091, F(3,315)=10.537, p<0.001			
Examiners	**0.25**	**4.25**	<0.001
PGCE students	**0.18**	**3.19**	0.002
Openness	**0.11**	**2.01**	0.045
Item 2b			

Variable	Beta	t	p
R^2=0.124, F(3,315)=14.850, p<0.001			
Examiners	**0.32**	**5.65**	**<0.001**
PGCE students	**0.18**	**3.27**	**0.001**
Enjoyment of marking	**0.11**	**1.98**	**0.048**
Part-script total **R^2=0.110, F(2,316)=19.430, p<0.001**			
Undergraduates	**-0.34**	**-5.98**	**<0.001**
English undergraduates	**-0.22**	**-3.83**	**<0.001**

Consensual view of 'true' score - correlation between the mean mark awarded by all participants and the marks awarded by individual participants to candidates' work

It is reassuring that employing a consensual rather than hierarchical measure of 'true' mark changes the conclusions of this analysis little (Table 12). Being a PGCE student is no longer positively associated with the extent to which items 2a and 2b are marked reliably but being an undergraduate is negatively associated with the marking reliability of these items, as is being an English undergraduate for item 2a. The role of judgement in applying the mark scheme proved to be an additional predictor of marking reliability at part-script level and for item 1c. Participants who believed that it is important to apply judgement in marking rather than strictly adhering to the mark scheme tended to mark less reliably. This is unsurprising since the mark scheme was the only information provided to guide their marking at this time.

Table 12 Independent predictors of the correlation between the mean of marks awarded by all participants and the marks awarded by individual participants to item and part-script responses

Variable	Beta	t	p
Item 1a **R^2=0.107, F(5,313)=7.532, p=<0.001**			
Examiners	**0.159**	**2.707**	**0.007**
Openness	**-0.203**	**-3.647**	**<0.001**
Agreeableness	**0.140**	**2.538**	**0.012**
Undergraduates	**-0.131**	**-2.267**	**0.024**
Gender	**0.112**	**2.055**	**0.041**
Item 1b R^2=0.082, F(2,316)=14.086, p<0.001			
Examiners	**0.180**	**3.163**	0.002

Variable	Beta	t	p
Undergraduates	**-0.173**	**-3.034**	0.003
Item 1c R^2=0.085, F(3,315)=9.807, p<0.001			
Undergraduates	**-0.198**	**-3.642**	<0.001
Role of judgement	**-0.147**	**-2.725**	0.007
Enjoyment of marking	**0.122**	**2.234**	0.026
Item 2a **R^2=0.171, F(3,315)=21.711, p<0.001**			
Undergraduates	**-0.332**	**-5.133**	**<0.001**
Examiners	**0.157**	**2.441**	**0.015**
English undergraduates	**-0.153**	**-2.336**	**0.020**
Item 2b **R^2=0.178, F(2,316)=34.120, p<0.001**			
Examiners	**0.340**	**6.323**	**<0.001**
Undergraduates	**-0.163**	**-3.028**	**0.003**
Part-script total **R^2=0.127, F(3,315)=15.294, p=<0.001**			
Undergraduates	**-0.336**	**-5.905**	**<0.001**
English undergraduates	**-0.208**	**-3.620**	**<0.001**
Role of judgement	**-0.113**	**-2.118**	**0.035**

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

Prior to training, marker background was a significant independent predictor of marking reliability (see Table 13). Examiners and PGCE students tended to mark item 1a and the part-scripts more reliably than participants with other backgrounds did. PGCE students' marking of items 1c and 2a was also more reliable than that of other participants. Both groups of undergraduates tended to mark item 1b less reliably than participants with other backgrounds did.

Participants' attitudes to marking were associated with how reliably they marked. The belief that only teachers should be employed to mark was an independent predictor of the difference in marks awarded for items 1b, 1c and the part-scripts. Participants with a relatively strong belief that only teachers should mark tended to mark less reliably. Perhaps this is a reflection of their feelings of self-efficacy. Participants who believed that it is important to apply judgement in

marking rather than strictly adhering to the mark scheme were also less likely to mark items 1c and 2a reliably. Those who reported enjoying marking were more likely to mark item 1c reliably.

Personality was associated with reliability of marking of items 1a and 2a. Participants who were open to new experiences were more likely to mark item 1a reliably. On the other hand, participants with relatively high scores on the agreeableness scale were less likely to mark item 2a reliably.

Age was associated with the differences in marks awarded at the level of part-script. Older participants were more likely to mark reliably than younger participants were. It is unclear what it is about being older that is independent of measures of personality, attitude and marker background, which makes older participants more likely to mark more reliably.

There were no significant, independent predictors of the difference in marks awarded to item 2b. Individual differences were unrelated to the absolute difference between the Principal Examiner's and participants' marking of this tem.

Table 13 Independent predictors of the absolute mark difference between the Principal Examiner's and participants' marking of item and part-script responses

Variable	Beta	t	p
Item 1a **R^2=0.137, F(3,94)=4.987, p=0.003**			
Examiners	**-0.33**	**-3.24**	**0.002**
Openness	**0.23**	**2.34**	**0.021**
PGCE Students	**-0.20**	**-2.08**	**0.040**
Item 1b **R^2=0.165, F(3,94)=6.181, p=0.001**			
English undergraduates	**0.51**	**4.14**	**0.000**
Undergraduates	**0.37**	**3.13**	**0.002**
Only teachers employed to mark	**0.32**	**2.85**	**0.005**
Item 1c R^2=0.254, F(4,93)=7.922, p<0.001			
Only teachers employed to mark	**0.28**	**3.06**	0.003
Enjoyment of marking	**-0.23**	**-2.55**	0.013
Role of judgement in marking	**0.28**	**3.07**	0.003
PGCE students	**-0.19**	**-2.06**	0.043
Item 2a R^2=0.154, F(3,94)=5.718, p=0.001			
Agreeableness	**-0.18**	**-1.81**	0.073

Variable	Beta	t	p
The role of judgement in marking	**0.24**	**2.40**	0.018
PGCE students	**-0.20**	**-2.02**	0.047
Part-script total **R^2=0.272, F(4,93)=8.705, p<0.001**			
Only teachers employed to mark	**0.53**	**4.20**	**0.000**
Examiners	**-1.12**	**-5.02**	**0.000**
Age	**0.62**	**3.10**	**0.003**
PGCE students	**-0.26**	**-2.86**	**0.005**

Consensual view of 'true' score - absolute difference between the mean mark awarded by all participants and the marks awarded by individual participants to candidates' work

There was more variation in the significant predictors of absolute mark difference than of correlation when a consensual rather than hierarchical measure of 'true' mark was used (Table 14). For example, taking the consensual 'true' mark for item 1a, participants who were relatively agreeable and participants who were relatively introvert were likely to have lower absolute mark differences. On the other hand, openness to experience was a significant independent predictor of absolute mark difference using the hierarchical measure of 'true' mark for this item. Although the pattern of predictors varied, the general conclusions are similar. In general, being an examiner or PGCE student was positively associated with marking reliability whereas being an undergraduate or English undergraduate was negatively associated with reliability. Strictly adhering to the mark scheme rather than applying judgement was associated with more reliable marking.

Table 14 Independent predictors of the absolute difference between the mean mark awarded by all participants and the marks awarded by individual participants to item and part-script responses

Variable	**Beta**	**t**	**p**
Item 1a R^2=0.100, F(2,101)=5.586, p=0.005			
Agreeableness	-0.323	-3.151	0.002
Extraversion	0.231	2.253	0.026
Item 1b R^2=0.105, F(2,101)=5.915, p=0.004			
Role of judgement	0.285	2.990	0.004
PGCE	-0.202	-2.118	0.037
Item 1c R^2=0.149, F(2,101)=8.819, p<0.001			

Variable	Beta	t	p
Role of judgement	0.294	3.193	0.002
Only teachers should mark	0.277	3.004	0.003
Item 2a R^2=0.125, F(2,101)=7.246, p=0.001			
Role of judgement	0.325	3.455	0.001
PGCE	-0.196	-2.086	0.039
Item 2b R^2=0.040, F(1,100)=4.118, p=0.045			
English Undergraduates	0.199	2.029	0.045
Part-script total R^2=0.081 F(1,100)=8.847, p=0.004			
Role of judgement	0.285	2.974	0.004

Summary of the predictors of pre-training marking reliability

The predictors of pre-training marking reliability varied depending on the item being analysed, on the operationalisation of 'true' mark and on the measure of reliability being used: correlation with Principal Examiner's marking or absolute distance of marks away from those awarded by the Principal. Moreover, only a relatively small amount of variance in marking reliability was measured by the variables. Taking an overview, examiners and PGCE students tended to be more reliable markers than undergraduates and English undergraduates. Hence, these findings confirm the earlier analyses of variance that explored the effect of background and training on marking reliability. The evidence suggests that in general prior to training PGCE students were as reliable markers as examiners.

Believing in strict adherence to the mark scheme was associated with higher reliability, as was enjoying marking and thinking that marking should be opened up to non-teachers. Tending to be relatively agreeable/co-operative was associated with marking that was more reliable, as was being conscientious. Findings regarding the importance of being open to new experiences were contradictory, being associated with higher reliability in some cases and lower in others. Generally, extraversion did not predict marking reliability but there is some evidence to suggest that introverts tended to mark more reliably than extraverts did.

The reliability of marking post-training

Principal Examiner's mark as the 'true' score - correlation between the marks awarded by the Principal Examiner and marks awarded by the participants

The following analyses focus on the correlation between the Principal Examiner's marking of the and that of the participant (Table 15). There were no significant, independent relationships between any of the predictor variables and the reliability of marking item 1a. This suggests that

this item can be marked equally well by anyone who has been trained, although it is possible that individual differences other than those measured in this study could be important. However, this item was marked out of three and the restricted range of scores will reduce statistical power of an analysis based on correlation. To investigate this possibility a logistic multiple regression was conducted. The reliability coefficient was recoded around the median to produce a dichotomous variable reflecting high and low reliability. The analysis produced two independent predictors of marking reliability (χ^2 = 9.134, df = 2, p = 0.010). Conscientious participants were more likely to mark this item reliably (beta = 0.032, Wald = 4.460, p = 0.035) and so were males (beta = 0.626, Wald = 5.830, p = 0.016). It is worth considering what it is about being male, rather than female, that is independent of measures of personality and attitude and marker background, that makes males more likely to mark this item reliably.

Moving on to item 1b, age was the only independent predictor of the extent to which this item was marked. Older participants tended to mark more reliably than younger participants did. Again, it is not immediately clear what it is about older participants, over and above their personality, attitude to marking, and marker background that leads them to mark this item more reliably. Moreover, extremely robust evidence of this age effect would be needed to support the active recruitment of older rather than younger examiners. Indeed the same argument applies to discrimination based on the association between gender and marking reliability.

Following training PGCE students were less reliable than other participants were in marking item 1c. It seems that the marker standardisation training was ineffective for PGCE students. Perhaps some aspect of their teacher training explains their response to the training. Participants with relatively high scores on the conscientiousness scale marked this item more reliably.

Conscientiousness was also an independent predictor of the reliability of item 2a marking. Further, examiners tended to be significantly more reliable in their marking of this item than other participants were and females' marking was more reliable than that of males. Examiners were also more able to mark item 2b reliably than participants with other backgrounds were. Participants with relatively high scores on the neuroticism scale were less reliable in marking this item. This interesting pattern of associations suggests that examiners, even neurotic ones, tended to mark this item reliably. However, if one were to employ non-examiners to mark this item, one would want to avoid those individuals with neurotic tendencies, a rather difficult recruitment policy to put into practice.

Conscientiousness was the only independent predictor of the extent to which the part-script as a whole was marked reliably such that a high score was associated with higher marking reliability. In systems where whole scripts are marked by examiners, rather than marking allocations being at an item level, it would seem that measuring individuals' conscientiousness would give an estimate of whether their marking is likely to correlate with that of the Principal Examiner.

Table 15 Independent predictors of the correlation between the Principal Examiner's and participants' marking of item and part-script responses

Variable	Beta	T	p
Item 1b **R^2=0.163, F(1,319)=8.710, p=0.003**			
Age	**0.163**	**2.951**	**0.003**
Item 1c R^2=0.063, F(2,318)=10.726, p<0.001			
PGCE students	**-0.223**	**-4.108**	<0.001
Conscientiousness	**0.122**	**2.251**	0.025
Item 2a R^2=0.102, F(3,317)=12.003, p<0.001			
Examiners	**0.221**	**4.031**	<0.001
Gender	**-0.141**	**-2.603**	0.010
Conscientiousness	**0.139**	**2.516**	0.012
Item 2b **R^2=0.065, F(2,318)=13.803, p<0.001**			
Examiners	**0.242**	**4.482**	**<0.001**
Neuroticism	**-0.123**	**-2.282**	**0.023**
Part-script total **R^2=0.029, F(1,319)=9.618, p=0.002**			
Conscientiousness	**0.171**	**3.101**	**0.002**

Consensual view of 'true' score - correlation between the mean mark awarded by all participants and the marks awarded by individual participants to candidates' work

The predictors of the size of the correlation changed when a consensual rather than hierarchical measure of 'true' mark was used (Table 16) but largely the conclusions were similar. Generally, being an examiner or PGCE student was positively associated with marking reliability. Agreeableness rather than conscientiousness was a significant predictor of marking reliability using the consensual 'true' mark. Participants with high scores on the agreeableness scale tended to mark less reliably. Using the hierarchical 'true' score, male participants tended to mark item 2a more reliably than female participants did. Using the consensual 'true' score, this was also the case for items 1b and at the level of part-script.

Table 16 Independent predictors of the correlation between the mean mark awarded by all participants and the marks awarded by individual participants to item and part-script responses

Variable	Beta	t	p
Item 1a **R^2=0.021, F(1,319)=6.735, p=0.010**			
Agreeableness	**0.144**	**2.595**	**0.010**
Item 1b R^2=0.057, F(2,318)=9.524, p<0.001			
Examiners	**0.183**	**3.353**	0.001
Gender	**-0.168**	**-3.066**	0.002
Item 1c R^2=0.082, F(2,318)=14.132, p<0.001			
Examiners	**0.241**	**4.399**	<0.001
Agreeableness	**0.112**	**2.047**	0.041
Item 2a **R^2=0.103, F(2,318)=18.209, p<0.001**			
Examiners	**0.294**	**5.524**	**<0.001**
Gender	**-0.154**	**-2.887**	**0.004**
Item 2b **R^2=0.117, F(2,318)=21.010, p<0.001**			
Examiners	**0.308**	**5.818**	**<0.001**
Neuroticism	**-0.120**	**-2.266**	**0.024**
Part-script total **R^2=0.056 F(2,318)=9.441, p=<0.001**			
Examiners	**0.205**	**3.758**	**<0.001**
Gender	**-0.136**	**-2.491**	**0.013**

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

The following analyses focus on the absolute difference in marks awarded to work by the Principal Examiner and the participants (see Table 17). Following training there were far fewer independent, significant predictors of mark differences. Once trained, participants' personal characteristics (attitudes, beliefs and personalities) had less impact on their ability to mark. Nonetheless, participant background remained influential. PGCE students were more reliable

markers of Item 2a than other participants were. On the other hand, both groups of undergraduates were less likely to mark item 2b reliably compared to other participants. Participants who believed that it is important to apply judgement to their marking rather than sticking to the mark scheme had larger absolute mark differences on item 2a than other participants.

Table 17 Independent predictors of the absolute mark difference between the Principal Examiner's and participants' marking of item and part-script responses

Variable	Beta	t	p
Item 2a R^2=0.092, F(2,95)=4.800, p=0.010			
The role of judgement in marking	**0.25**	**2.51**	0.014
PGCE students	**-0.21**	**-2.12**	0.037
Item 2b **R^2=0.110, F(2,95)=5.869, p=0.004**			
English undergraduates	**0.36**	**3.37**	**0.001**
Undergraduates	**0.22**	**2.06**	**0.042**

Consensual view of 'true' score - absolute difference between the mean mark awarded by all participants and the marks awarded by individual participants to candidates' work

When a consensual 'true' mark was used, the independent predictors of absolute mark difference changed somewhat (Table 18). In particular, being an English undergraduate was associated with lower marking reliability at the level of part-script. The belief that examiners should always have teaching experience was also positively related to marking reliability at the level of part-script.

Table 18 Independent predictors of the absolute difference between the mean mark awarded by all participants and the marks awarded by individual participants to item and part-script responses

Variable	**Beta**	**t**	**p**
Item 2a R^2=0.034, F(1,251)=8.845, p=0.003			
English Undergraduates	0.184	2.974	0.003
Item 2b R^2=0.053, F(2,249)=6.950, p=0.001			
English Undergraduates	0.249	3.675	<0.001
Undergraduates	0.143	2.107	0.036
Part-script total			

Variable	Beta	t	p
R^2=0.048 F(2,249)=6.209, p=0.002			
English Undergraduates	0.203	3.105	0.002
Only teachers should mark	0.169	2.585	0.010

Summary of the predictors of post-training marking reliability

As for the analyses relating to pre-training marking reliability, the predictors of reliability varied depending on the item being analysed, on the operationalisation of 'true' mark and on the measure of reliability being used (correlation or absolute mark difference). Taking an overview, examiners tended to be more reliable markers than both groups of undergraduates were. Prior to training, evidence suggested that PGCE students' marking was as reliable as that of the examiners. Following training, however, the evidence relating to the reliability of their marking is less clear. Earlier analyses suggested that the standardisation training was ineffective for this group of markers.

Once trained, participants' personal characteristics (attitudes, beliefs and personalities) were less likely to predict marking reliability. Nonetheless believing in strict adherence to the mark scheme was associated with higher reliability, as was thinking that marking should be opened up to non-teachers. Tending to be relatively agreeable was associated with better marking, as was being conscientious and emotionally stable (that is not neurotic). Extraversion and openness were not associated with marking reliability.

Discussion

It is necessary to draw some general conclusions that are not associated with particular operationalisations of marking reliability and that do not relate to specific items. Let us first consider whether individuals who are not experienced teachers could 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. Prior to training, there was little evidence of a significant difference in the marking reliability of examiners and PGCE students. Unfortunately, the marker standardisation training either failed to improve the reliability of the PGCE students' marking or even caused it to deteriorate. If PGCE students were to be employed as markers of longer answer questions they would require customised training. The qualitative evaluations of the current training gathered from the PGCE students did not highlight any specific problems. Indeed, they gave mostly positive evaluations but said that they would have liked more. 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.

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.

The analyses conducted and conclusions drawn have used the marking reliability of the examiners as a point of comparison (a gold standard). Both the groups of 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.

Next, let us consider whether psychometric measures of personality could be used as a tool for predicting those individuals likely to mark most reliably. Measures of neuroticism and extraversion were largely unrelated to marking reliability. There was some evidence to suggest that openness to experience was associated with marking reliability but only prior to training.

Individuals with high scores on this measure tend to have a positive attitude to learning new skills. However, it is necessary to predict how reliably individuals are likely to mark *following* training. Agreeableness (which can also be thought of as cooperativeness) was positively associated with marking reliability both before and after training, conscientiousness with reliability after training. These associations were independent of the other variables investigated in this study such as participants' backgrounds and attitudes. Although the relationships were relatively weak, accounting for only small amounts of variation in marking reliability, the difficulties of identifying any variables that consistently predict marking reliability (an inherently noisy variable) must be borne in mind. Before trialling the operational usefulness of these measures, the relationship between agreeableness and conscientiousness and marking reliability should be confirmed. This could be achieved at relatively low cost by having a sample of examiners involved in live marking complete a personality measure prior to marking.

Any attempt to use measures of attitude operationally in examiner recruitment and selection would be flawed since applicants would be able to 'fake good'. Moreover, participants' attitudes to marking predicted marking reliability prior to training but not following training. Training eradicated the impact of attitudes on marking reliability, surely a positive effect.

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 should be re-evaluated in the light of these findings.

There was some evidence that older participants tended to mark certain items more reliably than younger participants did. This raises the question as to what it is about older participants, over and above their personality, attitude to marking, and marker background that leads them to mark certain items more reliably. Perhaps it is a more general aspect of life experience that allows them to better judge responses to these questions. Perhaps older markers are more likely to view responses in the same way as the Principal Examiner who was himself middle-aged. Moreover, extremely robust evidence of this age effect would be needed to support the active recruitment of older rather than younger examiners. Equally, it is not immediately apparent why male participants marked some items more reliably than females and vice versa. Again, the evidence is not strong enough to support any discrimination based on gender.

The relationships between personality and demographic factors and marking reliability are clearly complex. We need to understand more about the characteristics of items that mediate these relationships before we will be able to predict who will be able to mark a particular item reliably and who will not. Surface characteristics such as the extent to which expert subject knowledge is required to mark the item do not seem to explain the links between reliability and personality and demographic factors. Moreover, it may be that the way in which the marker standardisation training was delivered accounts for some of these relationships. For instance, was there something about the training relating to item 2b that confounded participants with neurotic tendencies? The ephemeral nature of the training makes it difficult to know but this possibility will be investigated through discussion with the Principal Examiner.

Michelle Meadows and Lucy Billington, January 2007

References

Anderson, N. & Cunningham-Snell, A. (2000) Personnel selection. In N. Chmeil (Ed.) *Work and organizational psychology*. Oxford: Blackwell.

AQA. (2003) *Training for Senior Examiners. Reasons for mark compression and proposals for improving the spread of marks.* Research paper written and produced by staff at AQA to support Distance-Learning Pack 1.

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

Barrick, M.R., Mount, M.K. & Judge, T.A. (2001) Personality and performance at the beginning of the new millennium: what do we know and where do we go next? *International Journal of Selection and Assessment*, v9 p9-30.

Barrick, M.R. & Mount, M.K. (1991) The big five personality dimensions and job performance: A meta-analysis. *Personnel Psychology,* v44 n1 p1-26.

Branthwaite, A., Trueman, M. & Berrisford, T. (1981) Unreliability of marking: further evidence and a possible explanation. *Educational Review*, v33 n1 p41-46.

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

Buchanan, D. & Huczynski, A. (2004) *Organizational behaviour: An introductory text* (5th edition). London: Prentice Hall.

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

Costa, P. & McCrae, R. R. (1992a) Four ways five factors are basic. *Personality and Individual Differences*, v13 p653-665.

Costa, P. & McCrae, R. (1992b) *NEO PI-R professional manual*. Odessa, FL: Psychological Assessment Resources.

Costa, P.T., Jr.,& McCrae, R.R. (1992c) *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five Factor Inventory (NEO-FFI) Professional Manual*. Odessa, FL: Psychological Assessment Resources.

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

Digman, J.M., & Takemoto-Chock, N.K. (1981) Factors in the natural language of personality: Re-Analysis, comparisons, and interpretation of six major studies. *Multivariate Behavioural Research*, v 16 p149-170.

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.

Fowles, D. (2002) *Evaluation of an e-marking pilot in GCE Chemistry: Effects on marking and examiners' views.* AQA Research Report, RC/190.

Fowles, D. (2005) *How reliable is the marking in GCSE English?* AQA Research Report, RC/RPA06DEFRP020

Goldberg, J.R. (1993) The structure of phenotypic personality traits. *American Psychologist*, v48 p 26-34.

Goodstein, L.D. & Lanyon, R.I. (1999) Applications of personality assessment to the workplace: a review. *Journal of Business and Psychology*, v13 n3 p291-322.

Greatorex, J. & Bell, J.F. (2002a) *Does the gender of examiners influence their marking?* Paper presented at the Learning communities and assessment cultures: Connecting research with practice, University of Northumbria.

Greatorex, J. & Bell, J.F. (2002b) *What makes a senior examiner?* Paper presented at the British Educational Research Association, University of Exeter

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

Levy-Loboyer, C. (1994) Selection and assessment in Europe. In H.C. Triandis, M.D. Dunnette, & L. M. Hough (Eds) *Handbook of industrial and organisational psychology* (2nd edition). Palo Alto, CA: Consulting Psychologists Press.

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.

McCrae, R. R. & Costa, P. T. Jr. (1987) Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology*, v52 p 81-90.

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

Pal, S. K. (1986) Examiners' efficiency and the personality correlates. *Indian Educational Review*, v21 n1 p158-163.

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) (2005) *Code of practice 2005/6*. 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.

Salgado, J.F. (2003) Predicting job performance using FFM and non-FFM personality measures. *Journal of Occupational and Organizational Psychology*, v76 p323-346.

Salgado, J.F. (1997) The five factor model of personality and job performance in the European Community. *Journal of Applied Psychology*, v82 n1 p30-43.

Sanders, P. & Vanouzas, J.N. (1983) Socialization to learning. *Training and Development Journal*, v37 p14-21.

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.

Smith, G.M. (1967) Usefulness of peer ratings of personality in educational research. *Educational and Psychological Measurement*, v27 p967-984.

Takemoto, N.K. (1979) *The prediction of occupational choice from childhood and adolescent antecedents.* Unpublished masters thesis, University of Hawaii, Honolulu, HI.

Tett, R. P., Jackson, D. N. & Rothstein, M. (1991) Personality measures as predictors of performance: A meta-analytic review. *Personnel Psychology*, v 44 p703-742.

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.

Whetton, C. & Newton, P. (2002) *An evaluation of on-line marking*. Paper presented at the 28th International Association for Educational Assessment Conference, Hong Kong, September 2002.