CAN WE PREDICT WHO WILL BE A RELIABLE MARKER?

Michelle Meadows

The Assessment and Qualifications Alliance, Research and Policy Analysis Department have conducted a large number of studies that have attempted to identify factors which 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. Most of the work, however, is not publicly available. This paper provides a review of this research and internationally published studies relevant to predicting marker reliability. The relative importance of the following variables are discussed: examining experience; teaching experience; subject knowledge; senior examiners' ratings of examiners' previous performance; and examiner traits such as logical reasoning capacity and personality. Overall the conclusion of the review is that the criteria used by UK awarding bodies to select examiners (subject knowledge and teaching experience) are not empirically supported. The paper ends with the description of work that is currently being undertaken to examine whether psychometric measures of personality can be used to predict the marking reliability of individuals with distinctly different levels of examining experience, teaching experience and subject knowledge.

Background

In the UK, the selection of markers for national examination systems is largely a matter of custom and practice. The following criteria are used by the AQA, which are comparable to those used by other UK awarding bodies:

Examiners should have:

- Suitable academic qualifications (usually a relevant degree or equivalent)
- At least three terms' teaching experience which should be:
    - Recent - usually within the last three years depending on length of experience
    - Relevant – usually in schools or colleges, but may include university lecturing experience, teaching abroad (depending on where), or private tutoring. Experience of teaching AQA specifications is considered helpful, but not essential.
- Resident in the UK (normally)

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 has 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 which 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 here.

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 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. 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 reliability.

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 which resulted in the degree grade being altered by a whole degree class while competents had more changes to their marks but within the same degree classification.

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

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

So 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

[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.

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 can 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. It is likely that these are differentially important in marking performance.

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

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 consistently failed to find an 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 which influence the reliability of marking?

Examiner traits and marking performance

Some attempts to link personality traits with marking performance have been made. 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 unlikely.

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. A study that attempts to tease out the influence of the personality, background, attitude and motivation of the marker on the reliability of marking of different item types is now underway.

Current Research Project into Marker Selection

Background

The National Assessment Agency (NAA) is responsible for the supervision of the delivery and modernisation of UK GCSE and A level examinations. The NAA initiated this AQA project to begin the development of a marker selection instrument and investigate whether it improves the selection of markers in terms of reliability in their marking performance.

Psychometric tests have long been used for employee selection in industry. This study explores their utility 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.

The organisational psychology literature reports an association between certain psychometric measures and individuals' performance in jobs with demands similar to marking. These measures are likely to be useful predictors of marker performance. These psychometric tests fall into two broad categories:

1. measures of ability and aptitude;

   these have been shown generally to be good indicators of future job performance. In this study the aptitude of individuals with different education, teaching and examining backgrounds for marking was investigated. Participants initially marked scripts with only the mark scheme for guidance. They then received standardisation training before continuing to mark. The way in which the markers *responded* to this training was used as a predictor of future marking accuracy.

2. measures of personality;

   the way that a person performs in a job does not depend solely upon aptitude; personality characteristics also play an important part. Used in conjunction with other measures and assessments, personality measures can help predict future job performance. The Conscientiousness dimension of the five factor model of personality (as measured by the Neuroticism, Extraversion, and Openness Personality Inventory, or NEO-PI (McCrae and Costa, 1990; Costa and McCrae, 1992) is one such measure. Costa (1992) reported correlations between ratings of job performance and NEO-PI scores in a national sample of over 1,500 men and women. The strongest pattern of correlations was with Conscientiousness, which was related to the amount, quality and accuracy of work, and to overall judgements of competence. Five of the six Conscientiousness facets – Competence, Order, Dutifulness, Achievement, Striving, and Self-discipline - were related to superior performance ratings even after controlling for age, sex and years of education.

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

The extent to which measures of ability, 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 will also provide 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).

It is likely that the relationship between markers' ability, personality, motivation and attitude, and marking reliability will interact not only with their background, but with the kind of item being marked. 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.

Methodology

Four groups of one hundred[2] participants were recruited to mark the same one hundred GCSE English A[3], Higher tier, Paper 1, Section 1[4] scripts (Table 1). The groups were:

- Experienced GCSE English B markers (high subject knowledge and high teaching experience);
- PGCE English undergraduates (high subject knowledge and some teaching experience);
- English undergraduates (high subject knowledge and no teaching experience);
- Non-English undergraduates (low subject knowledge and no teaching experience).

Initially participants marked a first batch of 100 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 markers. Participants then marked another batch of 100 scripts. Participants completed a condensed version of the NEO-PI (240 items), known as the NEO-FFI (60 items) and bespoke measures of attitude and motivation. At the end

[2] To achieve statistical power greater than 0.80 in a multiple regression with an effect size (R^2) of 0.05 with ten predictor variables

[3] GCSE English was considered a suitable subject because: historically there is evidence of relative unreliability in marking e.g. adjustments applied to the marking; 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.

[4] To increase the variety of work marked by participants, they marked only one section of the question paper. The section included two questions: the first required two relatively short answers and one slightly longer answer; the second required one longer answer.

of marking, participants were canvassed on their overall experience of marking using a questionnaire.

Table 1 Groups of markers participating in the study

	subject knowledge	teaching experience
Experienced GCSE English B markers	high	high
PGCE English undergraduates	high	some
English undergraduates	high	none
Non-English undergraduates	low	none

Analysis

Analysis is being undertaken currently. Estimates of marking reliability are being calculated, at item, paper section and part-script level, for each group of participants. A number of definitions of marking reliability are being used. The absolute and relative difference between participants' marks and those of the Principal Examiner are being calculated. As well as this hierarchical definition of a candidates 'true score', a consensus approach is being explored. The mean score for an item or paper section is being calculated for all participants and for the Experienced GCSE English B markers. Participants' marks will then be compared to these mean scores. Differences in marking reliability across the groups and across item types will be tested using multivariate analysis of variance.

Analyses are being conducted to test whether the measures of response to training, personality, motivation and attitude predict marking reliability independent of marker background. Possible interactions between scores on these measures and background variables will also be examined. Multiple regressions are being used to conduct these investigations.

Finally, specialist software for qualitative data analysis (NVivo) is being used to analyse the data generated by the open-ended aspects of the questionnaire to inform on issues relating to training and levels of support required for the different types of marker.

Likely conclusions

These analyses will 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?

In other words, the study's findings will inform: the criteria used to select examiners; the kinds of items individuals with different traits, abilities and backgrounds are best able to mark reliably; the kind of support and training that would enable them to mark reliably. It is anticipated that the findings of the study will be available in the autumn, 2006.

References

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

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.

Costa, P.T. (1992) The Big 5: Personality and work. Paper presented ay the Sixth European Conference of Personality, Groningen, The Netherlands.

Costa, P.T., Jr.,& McCrae, R.R. (1992) *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.

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.

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

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. and Costa, P.T. (1990) *Personality in Adulthood*. New York, Guildford.
Paulhus, D.L. (1991). Measurement and control of response bias. In J.P. Robinson, P.R. Shaver and L.S. Wrightsman (Eds.) *Measures of Personality and Social Psychological Attitudes* (pp. 17-59). San Diego, CA: Academic Press.

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.

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

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

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.