IS TEACHING EXPERIENCE NECESSARY FOR RELIABLE MARKING?

Lucy Royal-Dawson and Jo-Anne Baird

SUMMARY

Although hundreds of thousands of markers are recruited internationally to mark examinations, little research has been conducted on the selection criteria that should be used. Many countries insist upon teaching experience for markers and this has frequently become embedded in the cultural expectations surrounding the tests. Shortages in markers for some of the UK's national examinations has led to non-teachers being hired to mark certain items and changes in technology have fostered this approach. For a national curriculum English examination taken at age 14, this study investigated whether teaching experience is a necessary selection criterion. Fifty seven markers with different backgrounds were trained in the normal manner and marked the same 98 students' work. By comparing the marking quality of graduates, teacher trainees, teachers and experienced markers, this study shows that teaching experience was not necessary for most of the examination questions. Small differences in inter-rater reliability estimates on the Shakespeare reading and writing tasks were found, such that non-teachers were less reliable than those with teaching experience. A model for selection of markers to mark different question types is proposed.

INTRODUCTION

Results from examinations are put to critical uses such as deciding a future academic or career direction, or evaluating the performance of teachers, schools, and even a country's education system. Their use at such important junctures necessitates that they are a fair reflection of test takers' ability. Many aspects of testing practices determine fairness, and this paper focuses on the reliability of marking, in particular where subjective judgement is required. This paper addresses inter-rater reliability, which is a measure of consistency in marking obtained over different occasions between different markers. Some marking systems use double marking, where scores from at least two markers contribute to an individual's overall mark, to counteract individual inconsistencies, but many rely on single marking. The latter are of particular interest because in some countries, such as the UK, it is becoming increasingly difficult to recruit enough markers to mark all papers once, and the expansion of the pool of markers beyond the traditional teaching profession is underway. The question arises then as to whether non-teachers are capable of marking to the same quality as teachers.

A more technical description of marking reliability in the context of this study follows, but first the task of marking and who can do it is explored. Subjective marking is the evaluation of a test taker's response to a test item according to a set of agreed criteria. Markers might be required to grade a response according to previously agreed standards or they might be required to allocate marks against marking criteria and the standard is set later in the light of the results from all test takers. To be able to score test papers, markers are oriented to the test and the allowed responses of the marking criteria, and then mark trial responses until they are able to mark accurately within given boundaries of acceptability.

Generally, markers are recruited from the teaching cadre of the qualification to be assessed. Already grounded in the curriculum focus, examination format and procedure, teaching professionals meet the requirements for the job. Examples of this come from Canada (EQAO, 2006), France (Journal Officiel de la République Française, 1993), New Zealand (NZQA, 2006), Scotland (SQA, 2006), South Australia (SSABSA, 2006), USA for the Advanced Placement, Graduate Record Examination and Scholastic Aptitude Test essay (ETS, 2006), and internationally for the International Baccalaureate (IBO, 2006). Yet, this begs the question of whether non-teachers could do as good a job. Examples of recruitment policies not stipulating teaching experience exist in the UK (England, Wales and Northern Ireland) and Queensland, Australia. Applicants with relevant subject knowledge are additionally recruited in the UK (NAA, 2006). Applicants to mark the Queensland Core Skills test (QCS), which is used to scale groups of students to contribute to calculations for ranking candidates for university selection, do not even have to demonstrate subject knowledge, but rather experience in 'criteria-and-standards-based-assessment'. They are then recruited depending on 'satisfactory achievement in recruitment tasks' (QSA, 2006). Potential markers for one paper are screened according to their performance on tests of 'verbal and quantitative areas … to suit the final make-up of the Short Response sub-paper' (Kempe, *personal communication*, 2006).

Very little research literature exists on how to select people who will prove to be able to mark consistently so QSA are pioneering a screening process. Studies that have attempted to link individual personal or psychological characteristics to marking quality are described by Meadows and Billington (2005) in their thorough review of marking reliability literature. None of the studies that linked personality traits, as measured by inventories or questionnaires, to quality of marking yielded results suggesting strong relationships in either direction. Furthermore, Pinot de Moira (2003) found the only personal characteristic of experienced markers linked to higher quality marking was the number of years of marking experience, which was confounded because reliable markers are

engaged year after year and poor markers are not, so quality marking and length of service as a marker are not mutually exclusive.

Perhaps subject knowledge is a sufficient condition for markers. Positive comparisons between the marking reliability of teachers and non-teachers with subject knowledge have been found in a handful of studies including one of an English as a foreign language test (Shohamy, Gordon and Kramer, 1992), an oral test for Japanese language tour guides (Brown, 1995), and an Occupational English test (Lumley, Lynch and McNamara, 1994). Comparable levels of marking reliability were found even when the condition of subject knowledge was removed in a study of the analytical writing component of the Graduate Management Admission Test (Powers and Kubota, 1998). Lay markers and experienced markers were found to reach similar levels of marking quality. Similar results were found in electronic marking studies using tests with closed items, where carefully selected items, one or two word answer responses, are presented to markers on-screen. Non-subject specialists in English or Maths marked items from the UK's Year 7 Progress Tests of English and Maths to an acceptable quality (Whetton and Newton, 2002), a finding echoed in both an experimental and a live use of e-marking piloted by the Assessment and Qualifications Alliance. Non-subject specialists reliably marked selected items in a GCE Chemistry unit under experimental conditions (Fowles, 2002) and generalist markers marked items from a live GCSE French listening paper with minimal error rates (Fowles and Adams, 2005).

Whilst the option to recruit non-teachers is attractive when several studies have shown advantages to employing them to mark and there is a shortage of markers, their acceptability amongst the test users needs to be assured. Many examination cultures depend on teachers marking longer responses to provide face validity to the tests' users. Yet, a hint of public mistrust in the UK of non-teachers marking was evident in the coverage of non-teachers marking components of GCSE Religious Studies in summer 2005 (*Times Educational Supplement*, 2005 and *The Guardian*, 2005). The awarding body's assurance that all markers were suitably qualified to mark the papers or items allocated to them had to be given to quell complaints. It is likely that the majority of test users are unaware of the highly technical nature of examination processing. The demonstration of transparent practices and evidence of the suitability of non-teachers to mark are needed to foster public confidence in the system. This study attempts to investigate the issue for the UK's Year 9 English national curriculum test by comparing the inter-rater reliability of teachers, trainee teachers and graduates of the same subject with experienced markers.

Traditionally, inter-rater reliability is expressed in terms of the inter-rater correlation coefficient between the markers under study and an expert marker for the same set of test papers. Effectively, this is a measure of the agreement in the rank ordering of the candidates. Inter-rater reliability estimates for tests consisting of open format items, such as essays, tend to be lower than for tests of closed items. There are no published standards for the test used in the study, so examples of estimates from similar tests are reported. Newton (1996), for example, demonstrated reliability estimates on re-marked papers in the UK's GCSE English at subject level between 0.81 and 0.95. Reliability estimates for the reading component alone, consisting of several single mark items, were higher (between 0.85 and 0.91), than for the writing elements, consisting of open format items awarded out of a higher mark (between 0.74 and 0.92). Essays require the markers to interpret the response which can lead to differences of opinion and thus lower inter-rater reliability measures. Tests of multiple choice questions can yield perfect inter-rater reliability measures because there is no room for interpretation. Another feature of inter-rater reliability discussed by Meadows and Billington is that it increases as the number of components within the overall test increases. For example, Murphy (1978) quoted inter-rater reliability estimates on two GCE O-level English Language components of 0.75 and 0.91 and a combined paper estimate of 0.90.

There are limitations to inter-rater reliability expressed as a correlation coefficient. It fails to indicate the underlying distribution of the correlated variables and is not user-friendly to those who use examination results (Shurnik and Nuttall, 1968). By ignoring the underlying distribution of the two variables, the measure becomes inflated (Coffman, 1971), and as a measure of rank ordering, it ignores systematic differences between markers (Lunz, Stahl and Wright, 1994). Baird and Mac's (1999) meta-analysis of reliability studies conducted by the Associated Examining Board in the early 1980s demonstrated that even near perfect reliability estimates of 0.98 were associated with up to 15% of the candidates not achieving the same grade. A reduction in reliability to 0.90, which is a high correlation, saw between 40% and 50% of candidates not receiving the same grade. Experienced examiners in Powers and Kubota's (1998) GMAT study yielded inter-rater reliability estimates between 0.79 and 0.96, but the level of agreement was at most 56% on essays marked out of six. In the UK, critics of national curriculum tests, which are reported as one of six levels, have challenged their accuracy by suggesting that up to 40% of the pupils receive a level classification higher or lower than they deserve (Wiliam, 2001). These findings imply the need for reporting of the agreement rate in the grades assigned to also account for reliability.

Another measure of marking quality is the degree of accuracy in marking. That is, the size of the difference between two examiners. Murphy (1978) showed average mark differences of about 5.7% in a blind marking study. He supports the use of the average absolute mark change as the simplest way of describing marker consistency (Murphy, 1982). In the current study, the three measures described above are calculated for the different groups of markers against an expert marker, the chief marker for the test used.

METHOD

Participants

a. Marker

Four groups of participants composed the sample (Table 1). English graduates were recruited for their subject knowledge, and lack of teaching and marking experience. Post-graduates from teaching degrees were recruited for their subject knowledge and their small amount of classroom experience gained on teaching practice. Teachers with three or more years' experience represented an ideal normally sought to mark the test. Only those with no external marking experience were recruited for the teachers group. Experienced markers who had marked the test to an acceptable quality before, but had not marked in the current year were recruited to act as the control group.

Table 1: Groups of participants

	BA degree in English	Post-graduate teaching degree	Teaching experience	Marking experience	No. in group
1. Graduates	Yes	No	None	None	17
2. Trainee teachers	Yes	Yes	A little	None	16
3. Teachers	Yes	Yes	At least three years	None	15
4. Experienced markers	Yes	Yes	At least three years	Yes	9

The target number of markers per group was 20 but it was not met because people reneged on their commitment, or because of a lack of willingness to take part. Experienced markers were also being sought for the final stages of live operational marking, which additionally hampered their recruitment. The timing of the study enabled the recruitment process to mirror live marking, with minor changes in contract and wording to suit the study.

b. Chief marker, supervisors and trainers

As for operational procedures, a single chief marker provided the true mark in the study. He was selected on the basis of having the most experience of marking the 2003 test amongst all other senior examining personnel. He marked all of the test papers used in the study so that his marks could be compared with those of all other markers. There is no higher authority who could mark more accurately, by the standard operational definition for this examination.

Seven experienced markers from 2003 were hired to take on an adapted team leader role for the markers. They were selected on the basis of having excelled in marking or having been team leaders in 2003. Two trainers were recruited to run the two training days for the markers.

c. Candidates

The test papers of 100 pupils were selected at random from a population of 36,810 papers. The sample was stratified to reflect the school types in the population. The papers were anonymised before they were photocopied. Two of them were excluded from the analysis due to poor printing quality, making the total 98.

Test materials

The test was the UK's national curriculum Key Stage 3 English test for 2003 taken by Year 9 pupils. It had been administered to pupils as three components: reading, writing and Shakespeare. The reading component assessed reading through 13 short answer or single word tasks with 5 marks being the highest score for any single item. The writing component consisted of a single longer writing task marked out of 30. The Shakespeare component had a writing task marked out of 20, and a reading task which required candidates to write a shorter piece marked out of 18.

Pupils receive three scores for the test: a reading paper score (reading component and Shakespeare reading task combined) out of 50; a writing paper score (writing component and Shakespeare writing task combined) out of 50; and a test score out of 100. The paper and test scores were further assigned to levels according to the threshold boundaries set for 2003.

Training and marking procedures

As far as possible all training procedures were identical to those used for operational marking. Markers received training materials adapted from operational marking before attending two days of training which covered marking the three components and administration. The team supervisors acted as table manager to their team of eight or nine markers. The markers followed up the training by marking copied papers and sending them to their supervisor for feedback.

The markers each received the photocopied test papers of the 98 pupils in the same order to mark at their own pace for completion to a set deadline. Their marking was standardised by a first sample check of the same 22 pupils. They received feedback in the form of marks and commentaries prepared by the chief marker on ten of the papers.

RESULTS

Inter-rater reliability estimates

Pearson product-moment correlation coefficients between the scores of the chief marker and each of the markers in the study were calculated for the overall test scores, the two paper scores and the two component and two tasks scores. Table 2 summarises the comparisons for the overall test scores. The differences in size of the mean reliability estimates are small, though the estimates appear to be more spread out in the less experienced groups. Similar tiny differences were observed for the paper

scores and the reading and writing components scores. As would be expected, the estimates for the reading paper and reading component were higher than for the writing scores.

Table 2: Summary of correlation coefficients between each marker and the chief marker for overall test scores

English test total	N	Mean	SD	Minimum	Maximum
Graduates	17	0.89	0.03	0.85	0.95
Trainee teachers	16	0.91	0.02	0.87	0.94
Teachers	15	0.92	0.02	0.89	0.96
Markers	9	0.92	0.01	0.90	0.94

Reliability estimates showed the greatest deviation between the groups in the two Shakespeare task scores (Table 3). Tests[1] for the homogeneity of variance were carried out using the *F*-test for the ratio of the larger to smaller variance for all pairs of groups for all papers, components and tasks (Howell, 1992). Homogeneity of variance was found in almost all comparisons, including the overall test estimates. Five instances of heterogeneity of variance were found in comparisons of the graduates with either the trainee teachers or the markers, with the estimates of the graduates being more spread.

Table 3: Summary of correlation coefficients between each marker and the chief marker for the two Shakespeare tasks

	N	Mean	SD	Minimum	Maximum
Shakespeare reading task					
Graduates	17	0.78	0.07	0.65	0.90
Trainee teachers	16	0.81	0.06	0.69	0.89
Teachers	15	0.85	0.03	0.77	0.89
Markers	9	0.86	0.02	0.82	0.89
Shakespeare writing task					
Graduates	17	0.74	0.07	0.53	0.84
Trainee teachers	16	0.77	0.05	0.66	0.84
Teachers	15	0.79	0.05	0.65	0.85
Markers	9	0.80	0.03	0.75	0.85

To test the hypothesis that the reliability estimates were no different in magnitude between the four groups, a one-way analysis of variance was carried out on the correlation coefficients using the Welch procedure to take account of the unequal sample sizes and the instances of heterogeneity of variance. No significant differences were found for the reading paper, the writing paper, nor the reading and writing components. This suggests the reliability estimates from the four groups for the papers and components were indistinguishable. Significant results were found at test level and for the two Shakespeare tasks meaning the size of the reliability estimates between the four groups were different.

To determine whether teaching experience was a contributing factor to the differences, *a priori* t-tests were carried out between the experienced markers, acting as the control group, and the other three groups. The results are summarised in 4. At test level, both the graduates and the trainee teachers were found to have significantly lower reliability estimates than the experienced markers, but not the teachers. This difference was also found on the Shakespeare reading task. These two results

[1] When correlation coefficients constitute the variable of interest, the transformation of r to r' is used where $r' = (0.5)\log_e |(1+r)| / |(1-r)|$ *to take account of the skewed distribution of r about ρ* (Howell, 1992).

indicate that at least three years of experience may have contributed to higher reliability estimates. Only the graduates were found to have significantly lower reliability estimates than the experienced markers on the Shakespeare writing task, suggesting that even a small amount of teaching experience contributed to higher reliability estimates for this task.

Table 4: Results of *a priori* t-tests to test for differences in reliability estimates

	Comparing the experienced markers with …			
		T	Df	*p*
Overall test	Graduates	-2.56	24	0.02
	Trainee teachers	-2.30	23	0.03
	Teachers	-0.10	22	0.92
Shakespeare reading task	Graduates	-3.28	24	<0.01
	Trainee teachers	-2.23	23	0.04
	Teachers	-1.06	22	0.30
Shakespeare writing task	Graduates	-3.01	23*	0.01
	Trainee teachers	-1.62	23	0.12
	Teachers	-0.74	22	0.46

*degrees of freedom adjusted to take unequal variances into account

To summarise the findings here, at least three years' teaching experience was found to be a contributing factor to significantly higher reliability estimates at test level and on the Shakespeare reading task. Having at least some teaching experience, as gained by the teacher trainees, was found to be a contributing factor to significantly higher reliability estimates on the Shakespeare writing task. A lack of teaching experience did not make any difference on the other aspects of the test, namely the reading and writing components. In other words, teaching experience counts for certain tasks, but not others.

Percentage of same Key Stage levels awarded

The second measure of marking reliability was the agreement rate for levels awarded by the markers and the chief marker. Pupils received three levels: one for the reading paper, one for the writing paper and one for the test overall. Reading and writing scores were awarded at four levels: 4, 5, 6 and 7. The test scores were awarded at six levels: N, 3, 4, 5, 6, and 7. Table 5 shows the percentage of agreements for the test, reading paper and writing paper. Across all markers, the agreement rate for the test was 61.22% and for agreements to within one level, the rate was 97.67%. For the reading paper, the agreement rate was 65.30% and within one level, 98.22%; and similarly for the writing paper, 50.22% and 94.16%.

Table 5: Percentage agreement of levels assigned by the four marker groups

Marker group	Test level	Reading level	Writing level
Graduates	58.69	64.64	48.28
Trainee teachers	61.45	66.71	50.99
Teachers	62.65	67.22	50.37
Markers	58.85	60.83	52.28
All markers	61.22	65.30	50.22

To test whether the agreement rates were the same for the four groups, analysis of variance was carried out on the number of agreements per group for each of the three levels. No significant differences were found. These results suggest that there was no difference between the groups with regard to their accuracy at assigning pupils the same levels as the chief marker. In other words, having teaching experience made no difference to a marker's ability to award the same level as the chief marker.

Even though the graduates had lower reliability estimates than the teachers on the test scores and the two Shakespeare tasks, they were equally accurate at assigning the pupils' levels. In other words, the graduates were less able than the other groups to rank order the pupils in the same order as the chief marker, but were equally able to assign the same level. Hence teaching experience did not make markers more accurate at assigning levels.

Marking accuracy

The third measure of marking reliability investigated was the difference between a marker's scores and those of the chief marker recorded as a non-negative quantity. A larger deviation away from the chief marker would be shown as a larger mark difference. They were calculated from the total scores on the test, papers, components and tasks. Table 6 shows the mean absolute mark differences for the test scores, the two components and the two Shakespeare tasks for each group. The paper score differences are not shown because they are a combination of the components and tasks, masking differences when the component and tasks scores are added together. So that the size of mark differences can be compared between components and tasks, the mean difference is also expressed as a percentage of the maximum mark for that aspect of the test. Unsurprisingly, the written tests have largest mean differences for all groups.

Table 6: Mean and standard deviation of absolute mark differences for overall test scores

		N	Mean	SD	Mean expressed as % of maximum mark
English test total (out of 100 marks)	Graduates	1630	7.01	5.59	7.01
	Teacher trainees	1507	6.48	5.30	6.48
	Teachers	1447	6.09	5.07	6.09
	Markers	863	6.93	5.18	6.93
Reading component (32 marks)	Graduates	1639	2.33	1.94	7.28
	Teacher trainees	1536	2.43	2.03	7.59
	Teachers	1451	2.22	1.90	6.94
	Markers	863	2.63	2.20	8.22
Writing component (30 marks)	Graduates	1698	3.76	3.10	12.53
	Teacher trainees	1599	3.76	3.10	12.53
	Teachers	1498	3.62	2.99	12.07
	Markers	894	3.59	3.00	11.97
Shakespeare reading task (18 marks)	Graduates	1693	2.08	1.72	11.56
	Teacher trainees	1570	1.83	1.49	10.17
	Teachers	1498	1.73	1.47	9.61
	Markers	894	1.72	1.55	9.56
Shakespeare writing task (20 marks)	Graduates	1697	2.89	2.25	14.45
	Teacher trainees	1598	2.73	2.06	13.65
	Teachers	1499	2.73	2.25	13.65
	Markers	894	2.46	1.91	12.30

Tests for homogeneity of variance indicated that the absolute mark differences were similarly distributed for all groups only on the writing component. Instances of heterogeneity of variance were found in all other aspects of the test for most, though not all, of the comparisons. The pattern of differences in variability is not uniform and does not suggest one group is consistently more variable than the others on any or all aspects of the test.

To test for differences between the size of the groups' absolute mark differences, repeated measures analyses of variance were conducted. The pupils' absolute mark differences were the 98 within-subject variables and the four groups were the between-subject factor. In this procedure, the entire variation is partitioned so that the within-subject variation and interaction term are separated from the between-subject variation, and they have their independent error terms. For the purposes of the study, the within-subject analysis, including the interaction terms, was of no interest and only the between-subject effects were investigated. The between-subject procedure is robust against some violations of the assumptions (Howell, 1992), so the lack of homogeneity of variance between some groups was not a hindrance to using it. No significant between-subject effects were found. This suggests that the accuracy of the groups was at a statistically similar level on all aspects of the test. Thus, when compared with the chief marker no differences between the groups emerged. There were more and less accurate markers in each group with no group emerging as less accurate than any other. Again, teaching experience did not emerge as a distinguishing factor between the markers.

DISCUSSION

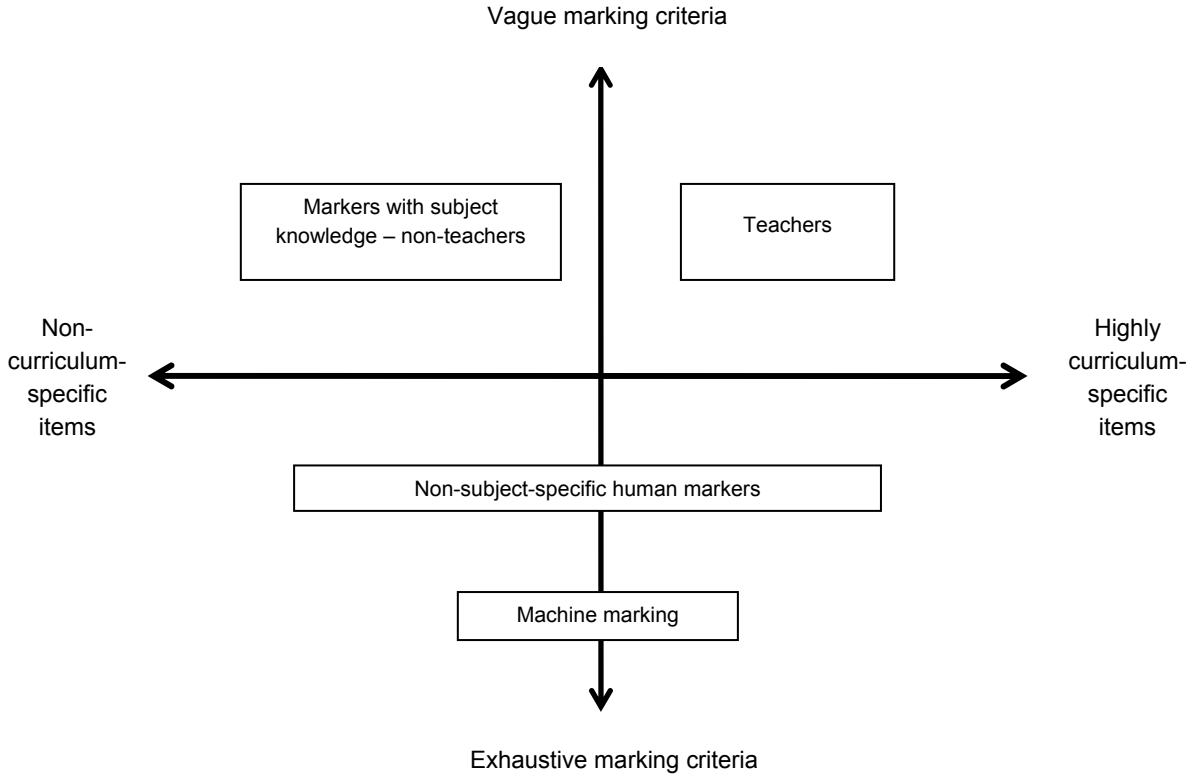
The markers in the study were capable of comparable levels of marking quality for the UK's Year 9 national curriculum English test with regards to mark differences and assigning levels. Teaching experience only made a difference when quality was measured by the inter-rater correlation coefficient, and that was only for the two Shakespeare tasks, an effect which transferred to the overall test scores. In other words, the marks awarded for the reading and writing components were of equal accuracy from graduates, teacher trainees, teachers of at least 3 years and experienced markers.

One interpretation of the findings for the Shakespeare tasks was that the teachers had some un-quantified advantage over the trainees and graduates. Perhaps the Shakespeare tasks required a familiarity with this type of exercise gained from classroom practice, or an idea of what to expect of students at this stage of their education. This suggests the degree to which an item is specific to a taught curriculum determines the professional requirements of markers. If the Shakespeare items had demanded a demonstration of general knowledge of Shakespeare's work, perhaps the graduates would have marked them more reliably. Curriculum-specific here means the degree to which an item can only be answered if a candidate has followed a particular course of study.

If teachers are needed to mark curriculum-specific items and less specific items can be marked by subject specialists, and we have already seen that people with no subject knowledge can mark closed items, a cross-classification of marker with item type emerges. Closed response formats have strict, containable marking criteria making their interpretation straightforward. Open items demand more complicated marking criteria depending on either the subject content or learning objectives to be demonstrated or both. Essays eliciting knowledge or comprehension only can be marked against exhaustive and containable marking criteria, whereas essays demanding demonstration of higher learning objectives require marking criteria that are less easily containable. Encapsulating all possible permissible answers in such marking criteria would be unfeasible and they will be, by necessity, vague. Figure 1 matches teachers, subject specialists, non-subject specialists and machines according to the nature of the items to be marked. This basic model is proposed for any subject content. Items with easily definable marking criteria, irrespective of how curriculum-specific they are, can be marked by machines or trained humans. Teachers and subject specialists are required to

mark items with marking criteria that are not easily contained. Teachers distinguish themselves from non-teaching subject specialists at the point where items are heavily dependent on the curriculum. This could account for graduates and teacher trainees marking the Shakespeare tasks less reliably. They were simply less well-versed in the curriculum and classroom.

Figure 1: Item type matching model for marker selection



One drawback to this model is the difficulty in identifying items which are curriculum-specific and therefore to whom they should be directed for marking. Fowles and Adams (2005) successfully allocated items from a GCSE French listening test between expert markers, generalists and machines, suggesting that categorisation of items is possible, but the distinction between more or less curriculum-specific items may only emerge once marking quality of different types of markers is compared.

An alternative explanation of the findings is that the experienced markers were not viable as a control group, producing no differences between the groups. Their past marking indicated they were highly regarded and would have met selection criteria to conduct live marking in 2003. That they did not mark earlier in the year should not account for much. People often take a break from marking and they do not have to undergo special re-induction procedures when they start again. The training in the study was the same as that used in the live marking, conducted by the same trainers in one of the same locations. Experienced markers reported that their main motivations for taking part were to get experience in the new test format of Key Stage 3 English and to make up a short-fall in income lost through not taking part in the live marking earlier in the year, suggesting motivation based upon a serious professional interest. There is no overt reason why they would not be representative of experienced markers.

The differences between the groups' variance measures of the absolute mark differences were not statistically significant, suggesting there were more and less accurate markers in each of the groups, but no group had more or fewer accurate markers than any other. It is interesting that the groups' mean absolute mark differences on the overall test were between 6.1% and 7.0% compared to 5.7% found by Murphy (1978), suggesting the mark differences were comparable in size to those found in other studies of experienced markers.

With regards to level agreement rate between the groups and the chief marker, it is interesting that the proportion of disagreements was roughly 40% as estimated by Wiliam (2001), and it mirrors the disagreement rate amongst experienced readers that Powers and Kubota (1998) saw in their study which similarly used a six point scale.

Reference to the literature (Murphy, 1978 and Newton, 1996) suggests that the reliability estimates at all levels of the test were on a par with those found in studies of other examinations, but it is difficult to assess what constitutes an acceptable level of marking reliability for this test, particularly with its intended purpose in mind (Newton, 2003). Wiliam (2003) argued that where results of tests are used in schools for formative purposes, such as this one, their reliability should be measured from the accuracy of decisions made based on evidence from test results in a pupil's school career. This would be a measure of the tests' consequential validity and it serves as a good reminder of the critical uses to which test results in general are put even in the absence of agreed and stated measures of marking reliability. The findings in the study may provide a basis for deriving a consensus of an acceptable degree of marking reliability for the Key Stage 3 English test. The reporting of such measures is recommended in *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1999), but is a rare practice in the UK.

A final point that needs to be considered is that of a study effect. The markers knew they were participating in a study and the marks they assigned would have no impact on the pupils, only on the study's outcomes. It is possible they were not as engaged with the task as they would be in live marking. The high degree of correlation between markers' and the chief marker's scores is evidence that the former appreciated the differences between pupils' performance and were not assigning marks with abandon. Also, the relatively small mark differences suggest that they applied themselves to the task. For these reasons, it is argued that a study effect was unlikely.

CONCLUSIONS

Marking does not have to remain the preserve of teachers of the curriculum being assessed. Building public confidence through transparency with the demonstration of evidence could open the way to allocating papers to markers with sufficient skills and knowledge. A cadre of professional markers put to work on any subject matter may yet come into existence. Items which are highly curriculum-specific may need to remain the teacher's lot, unless students, who are also immersed in the curriculum, mark their own or their peers' work: possibly one step beyond the public's tolerance in many current educational cultures.

DISCLAIMER

This work was funded by the National Assessment Agency in London. Opinions expressed in the paper are those of the authors and should not necessarily be taken to represent those of the National Assessment Agency or the Assessment and Qualifications Alliance. The design of the Key Stage 3 English tests has changed since this study was carried out and the figures quoted for reliability may not generalise to current tests.

REFERENCES

American Educational Research Association (AERA), American Psychological Association (APA) and National Council on Measurement in Education (NCME) (1999) *Standards in educational and psychological testing.* Washington DC: American Educational Research Association.

Baird, J. and Mac, Q. (1999) *How should examiner adjustments be calculated? A discussion paper. AQA* Research Committee paper, RC/13.

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

Coffman, W.E. (1971) *Essay Examinations.* In R.L. Thorndike (Ed.) *Educational Measurement,* Washington DC: American Council on Education.

EQAO - Education Quality and Accountability Office, Ontario (2006) *Opportunities for Ontario Educators Scorer Job Description*, accessed on 23rd March 2006 from http://www.eqao.com/Employment/Employment.aspx?Lang=E#scoring.

ETS - Educational Testing Services (2006) *Paper and Pencil Scoring Opportunities*, accessed on 23rd March 2006 from http://www.ets.org/portal/site/ets/menuitem.22f30af61d34e9c39a77b13bc3921509/?vgnextoid=88067f95494f4010VgnVCM10000022f95190RCRD.

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

Fowles, D. and Adams, C. (2005) *How does marking differ when emarking?* Paper presented at the 31st International Association for Educational Assessment conference, Nigeria, September.

The Guardian (2005) *Exam board under fire over marking claims (26.08.05),* accessed on 21 April 2006 from http://education.guardian.co.uk/gcses/story/0,,1557161,00.html.

Howell, D. (1992) *Statistical Methods for Psychology* 3rd ed. Belmont, California: Duxbury Press

IBO - International Baccalaureate Organisation (2006) *General Terms of Contract for IB Diploma Programme Assistant Examiners*, accessed on 23rd March 2006 from http://www.ibo.org/examiners/assistant_posts/assistantexaminers/documents/AEcontract1Aug02.doc.

Journal Officiel de la République Française, edition lois et décrets (1993) *Décret portant règlement général du baccalauréat general, Décret no. 93-1091 du 15 septembre 1993*, accessed on 23rd March 2006 from http://www.ac-nancy-metz.fr/enseign/lettres/inspection/Divers/baccalaureat_regl.htm.

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

Lunz, M.E., Stahl, J.A. and Wright, B.D. (1994) Interjudge reliability and decision reproducibility. *Educational and Psychological Measurement*, v54, p913-925.

Meadows, M.L. and Billington, L. (2005) *A Review of the Literature on Marking Reliability*. AQA Research Committee paper, RC304.

Murphy, R.J.L. (1978) Reliability of Marking in Eight GCE Examinations. *British Journal of Educational Psychology*, v48, p196-200.

Murphy, R.J.L. (1982) A further report of investigations into the reliability of marking of GCE examinations. *British Journal of Educational Psychology*, v52, n1, p58-63.

NAA - National Assessment Agency (2006) *Behind Every Paper There's A Student: Who can do it*, accessed on 23rd March 2006 from http://www.examinerrecruitment.org/whocandoit/default.htm.

Newton, P.E. (1996) The Reliability of Marking of General Certificate of Secondary Education Scripts: mathematics and English. *British Educational Research Journal*, v22, n4, p405-420.

Newton, P. (2003) The defensibility of national curriculum assessment in England. *Research Papers in Education,* v18, n2, 101-127.

NZQA – New Zealand Qualifications Authority (2006) *Role Definition: Marker 2005,* accessed on 28th March 2006 from http://www.nzqa.govt.nz/about/jobs/contracts/docs/marker.doc.

Pinot de Moira, A. (2003) *Examiner Backgrounds and the Effect on Marking Reliability.* AQA Research Committee Paper, RC/218.

Powers, D. and Kubota, M. (1998) *Qualifying Essay Readers for an Online Scoring Network (OSN).* Educational Testing Service, USA.

QSA - Queensland Studies Authority (2006) *Queensland Core Skills Test New Applicant Information Package*, accessed on 23rd March 2006 from http://www.qsa.qld.edu.au/testing/markers/docs/NewMarkerpackage2006.pdf.

Shohamy, E., Gordon, C. and Kramer, R. (1992) The effects of raters' background and training on the reliability of direct writing tests. *Modern Language Journal*, v7, n1, p27-33.

Shurnik, L.S. and Nuttall, D.L. (1968) Describing the reliability of examinations. *The Statistician*, v18, p119-128.

SQA – Scottish Qualifications Authority (2006) *Leaflet: You've Taken Them This Far*, accessed on 23rd March 2006 from http://www.sqa.org.uk/files_ccc/MarkersRecruitmentForm.PDF.

SSABSA - Senior Secondary Assessment Board of South Australia (2006) *Information on SSABSA Stage 2 Assessment Panels*, accessed on 23rd March 2006 from http://www.ssabsa.sa.edu.au/panels/marking.htm.

Times Educational Supplement (2005) *Markers had 20 minutes' training (26.08.05),* accessed on 21st April 2006 from http://www.tes.co.uk/search/story/?story_id=2126566.

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

Wiliam, D. (2001) *Level Best? Levels of Attainment in National Curriculum Assessment*. Association of Teachers and Lecturers, UK

Wiliam, D. (2003) National curriculum assessment: how to make it better. *Research Papers in Education*, v18, n2, 129-136.