Assessment and Qualifications Alliance

IS TEACHING EXPERIENCE A NECESSARY CONDITION FOR MARKERS OF KEY STAGE 3 ENGLISH?

Lucy Royal-Dawson

Commissioned by the Qualifications and Curriculum Authority.
QCA, 83 Piccadilly, London, W1J 8QA. United Kingdom.

IS TEACHING EXPERIENCE A NECESSARY CONDITION FOR MARKERS OF KEY STAGE 3 ENGLISH?

Report of the Key Stage 3 English Marker Study

EXECUTIVE SUMMARY

1. One of the main criteria for recruiting examiners is the number of years of experience as a classroom teacher. The difficulty of recruiting sufficient numbers of markers to Key Stage 3 English prompted a formal study of whether this criterion could be relaxed. Four groups of markers representing markers with distinctly different teaching and marking backgrounds were investigated. They were experienced Key Stage 3 English markers, experienced English teachers with no external marking experience, recent PGCE graduates of English with teaching experience gained on teaching practice and recent BA graduates of English with no teaching experience.
2. The markers were recruited and trained using the same procedures and documentation as that used in live marking. They marked the same allocation of 100 photocopied scripts. Their marks were analysed and compared to those given by the highest authority on Key Stage 3 English marking, the Lead Chief Marker.
3. The findings indicate no overwhelming differences between the groups, only a number of minor ones. The main findings are:
    a. There was no difference in marking accuracy between the four groups. When compared to the Lead Chief Marker, all groups achieved similar levels of accuracy, suggesting that teaching experience was not a contributing factor.
    b. There was no difference in marking reliability between the four groups as defined by the agreement rate of Key Stage levels assigned by markers compared to the Lead Chief Marker's. This suggests that teaching experience was not a contributing factor. The combined agreement rate across all four groups was 61% and the rate accurate within one level was 98%. No precedence for accuracy in Key Stage 3 English exists in the literature.
    c. Marking reliability as defined by the correlation coefficient between each marker and the Lead Chief Marker indicated that some teaching experience was a contributing factor to higher reliability estimates in the two Shakespeare tasks and the overall test. The reading and writing components did not reveal any differences between the groups, suggesting that teaching experience was a contributing factor on some tasks but not on others.
    d. 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.
    e. The relationship between the two measures of marking reliability, the agreement rate in level assignation and the correlation coefficients, was found to be not clearly defined so that a high correlation coefficient was not necessarily associated with a high agreement rate. The author questions the efficacy of measures of correlation as a means of assessing marking reliability.

   f. The method of determining the acceptability of markers at first phase sampling was shown to be potentially flawed when it relies on absolute mark differences alone. An exploration of the hypothetical rejection of some markers on the basis of high absolute mark differences suggested some markers who were capable of acceptable marking would have been prematurely rejected.

4. The study concludes that the criterion of teaching experience could be relaxed to allow markers with graduate-level subject knowledge to mark Key Stage 3 English tests. Whether other Key Stage levels and subjects could be marked by non-teachers should also be explored. The use of absolute mark differences as a means of assessing markers and the relationship between traditional measures of reliability should also be further investigated. The rate of disagreement in level assignation is a cause for concern when there is no procedure for marker adjustment. This is especially so since much importance is attached to Key Stage 3 results. Further investigation of the data would be necessary to establish whether marker adjustments could have increased the agreement rate.

1. BACKGROUND

One of the main criteria used by Awarding Bodies for evaluating the employability of an examiner or marker[1] is relevant classroom experience. Generally, examiners are teachers with a number of years of teaching experience that is deemed sufficient to enable them to make good judgements when marking scripts. This study investigates whether the criterion of classroom experience can be removed.

The recruitment of markers for Key Stage 3 English to numbers sufficient for total coverage is increasingly becoming difficult. A total of 1,405 markers were recruited to mark 2003's scripts. The use of markers who do not fit the current marker recruitment criteria could increase the pool of markers and thus ease the shortage. The difficulty of doing this arises in determining who would make suitable Key Stage 3 English markers and what marking reliability and accuracy is required of them. More specifically, these areas of uncertainty were framed by the five following questions:

1. Can English graduates and PGCE graduates be considered as potential markers?
2. What levels of marking reliability can different types of markers achieve compared to experienced markers?
3. Is the accuracy of marking by different types of markers acceptable for live marking?
4. What training, standardisation and support requirements would different types of markers have?
5. What criteria for recruitment and conditions of service would be appropriate in the employment of different types of markers?

The first three questions above relate to recruitment criteria for markers and the quality of marking different markers can achieve. In order to investigate further, the Qualifications and Curriculum Authority provided funds for a marker comparison study (Royal-Dawson, 2003). The last two questions relate to the practical implications of recruiting new types of markers. The study addresses the first three questions directly and raises issues associated with the last two.

The first question relates to the employment of markers who do not fit the usual criteria. The criteria for selecting Key Stage 3 English markers are specified in terms of a hierarchy of preferred attributes. At the top:

> Qualified teacher with at least 3 years secondary experience, currently teaching the appropriate subject at key stage 3 on a full or part time basis.

People fulfilling this criterion prove hard to recruit and instead people are selected who meet criteria lower down the list:

[1] The term 'examiner' is used to denote marking personnel in general and the term 'marker' is used to denote national curriculum marking personnel specifically.

Qualified teacher with relevant teaching experience within the last three years and of at least three years overall including:

- Serving headteacher or deputy headteacher;
- Advisor, inspector, LEA literacy/numeracy/ strategy consultant;
- Supply teacher;
- Retired teacher.

or

Qualified teacher with relevant teaching experience not within the last three years, but of at least three years overall who is a:

- Serving headteacher or deputy headteacher;
- Advisor or inspector.

or

Qualified teacher, currently teaching full or part time in their second year of teaching but with less than three years relevant experience.

or

Retired and qualified teacher with relevant teaching experience not within the last three years, but of at least three years overall.

The common factor in these criteria is three years of teaching experience. A marker must be a qualified teacher. Studies investigating the viability of electronic marking that have attempted to query the need for teaching experience have had some success in showing that non-teaching professionals can be trained to mark to an acceptable quality. Powers and Kubota (1998) showed that current requisites for essay markers for the Graduate Management Admission Test (GMAT) would disqualify people who could be trained to mark, though they draw the line at saying that anyone could be trained to mark. They removed the requisite that markers should be currently teaching on a graduate course and found that many of the inexperienced markers in the study marked to the same level of accuracy as the markers meeting the requisites. They even suggested that pre-screening markers might reveal the people who would go on to be accepted for marking. The use of unskilled or semi-skilled markers for carefully selected items of a Year 7 Progress Test in a study investigating the use of electronic marking was also deemed successful (Whetton and Newton, 2002). This finding was echoed in an AQA e-marking study, where clerical markers reliably marked selected items in a GCE Chemistry unit (Fowles, 2002). In Pinot de Moira's (2003) investigation of the effects of examiner backgrounds on marking reliability, the only personal characteristic found to be significant in explaining examiner reliability was the number of years of marking experience. Yet, this characteristic 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. Teaching experience was not investigated because it was a characteristic of all the markers.

The present study tackles the requisite of teaching experience by comparing the four groups of markers. Unlike those in the Powers and Kubota (1998) study, all markers had an academic background in the subject, English. The groups are:

1. BA graduates of English in 2003,
2. PGCE graduates of Secondary English in 2003,
3. teachers with three or more years' teaching experience, that is, who meet the first criterion, and
4. experienced markers who have marked Key Stage 3 English before.

The rationale for selecting these four groups and the criteria used to recruit them are given in Section 2 Methodology. The marks they assigned to the same sample of scripts were compared for differences that might suggest teaching experience sets apart acceptable markers. The results of the comparisons are reported here.

The second question posed in the study relates to what constitutes reliable marking. For a marker to be acceptable he or she must reach a minimum standard of marking reliability. This in turn suggests that currently employed markers reach a level of reliability that is acceptable. The training, standardisation and sampling procedures that markers must follow have built-in gauges of marking acceptability. Markers are judged as suitable to continue marking if they reach pre-determined levels of accuracy on their sample scripts. Yet, critics of national curriculum tests have challenged their accuracy tests by suggesting that the unreliability of the tests results in up to 40% of the pupils receiving a level classification higher or lower than they deserve (Wiliam, 2001). He argues that since the results of national curriculum tests may be used in schools for formative purposes, the reliability of Key Stage tests should be a measure of the accuracy of decisions made based on evidence from the tests in a pupil's school career (Wiliam, 2003). Unpicking reliability into the separate threads of test-retest reliability, mark-remark reliability and level classification, Newton (2003) responded to Wiliam's accusation by: noting that test-retest reliability estimates were not available for some papers of Key Stage 3 English and that the estimates achieved on other papers were reasonable; accepting that the number of successful requests for remarking suggested that 'national curriculum tests are not perfectly marked' and noting that there have been no published studies of marking reliability for any of the tests; and arguing that even tests with very high reliability statistics will result in some misclassifications, and that it is necessary to establish a consensus on the level of acceptable reliability for the tests' intended purpose. This study goes some way to explore the marking reliability of the Key Stage 3 English tests.

In the absence of Key Stage 3 English marker reliability literature, studies of other specifications are referenced. Traditionally, marker reliability is expressed in terms of the inter-rater correlation coefficient between the marks of a marker under study and a principal marker from the same set of scripts. This is a measure of the agreement of the rank ordering of the candidates. Murphy (1978) quoted marking reliability estimates on scripts with marks and comments removed for two GCE O-level English Language papers at 0.75 and 0.91 and a combined paper estimate of 0.90. In a study where marks were left on some essay scripts and taken off others, differences in marking reliability were revealed (Murphy, 1979). 'Marks on' scripts yielded reliability estimates between 0.94 and 0.96, while 'marks off' scripts yielded

estimates between 0.85 and 0.87. Newton (1996) demonstrated reliability estimates on remarked 'marks off' scripts in GCSE English at subject level between 0.81 and 0.95. Breaking the total marks into constituent elements of the mark scheme, he found higher reliability estimates for the reading element which consisted of several single mark items, between 0.85 and 0.91, than for the writing elements which consisted of free-response items awarded out of a higher mark, between 0.74 and 0.92. The present study reports the marking reliability of four types of markers at test and component levels.

The usefulness of reliability estimates expressed as correlation coefficients is limited if the proportion of candidates receiving the same grade or level is not also quoted. Baird and Mac (1999) conducted a meta-analysis of reliability studies conducted by the Associated Examining Board in the early 1980s to show the relationship between inter-rater reliability measures and the proportion of candidates getting the same grade. They 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 still a reasonable estimate, saw between 40% and 50% of candidates not receiving the same grade. Experienced markers in Powers and Kubota's (1998) study yielded reliability estimates between 0.79 and 0.96, but their level of agreement was at most 56% on essays marked out of 6. There are six levels in Key Stage 3 English, which suggests that Wiliam's (2001) criticism, that they yield the wrong level for up to 40% of pupils, may be plausible. The present study will provide an indication of misclassifications for each marker group and it will report on them as a way of shedding light on the reliability estimates.

The third question raised by the study relates to the level of accuracy that is acceptable for live marking. The size of the difference between two markers is used as a means of investigating accuracy, in this case between the Lead Chief Marker for Key Stage 3 English and each marker. Murphy (1978) showed average mark differences of about 2.5% on 'marks on' scripts and about 5.7% on 'marks off' scripts. Meadows (*in preparation*) investigated the size of mark difference between examiners and team leaders at the first standardisation sample (that is, a sample of scripts marked by the examiner and sent to his or her team leader for re-marking as a means of assuring accurate marking) across a range of GCE units. Both 'marks on' and 'marks off' scripts were used in the samples. For example, mark differences for English Literature 'marks on' scripts ranged from 13.9% to 31.0% and from 14.5% to 46.6% for 'marks off' scripts. The mark differences from the different types of markers are explored in the present study to investigate whether teaching experience is a contributing factor to smaller mark differences.

The findings related to the use of 'marks on' and 'marks off' scripts are particularly pertinent to the current study because the procedure for the first sample, part of the on-going marker standardisation process, differed to the live marking procedure. Live procedures involve team leaders re-marking a sample of their markers' scripts and giving feedback to the markers on any differences. This corresponds to the 'marks on' scenario where the team leader can see the marker's mark. Murphy (1978) suggested that

> examiners who are asked to re-mark scripts cannot help but be influenced by these previous judgements, however much they try to ignore them and form their own opinion.

Massey and Foulkes (1994) argue that being able to see how the marker came to a decision provides insight to the team leader that enables him or her to support the original decision. In Key Stage 3 English in 2003, a difference between the team leader and marker's marks of 66 marks or more on the ten scripts, which translates to 6.6% of the total marks, was deemed a cause for concern (AQA, 2003). However, because the study coincided with live marking review, the team leaders were not available for this duty and instead a 'marks off' scenario was used for the first sample. The markers all marked the same sample scripts and instead of their marks being mediated by different team leaders, they received the Lead Chief Marker's marks for the ten with a commentary on how the marks were awarded. Since both Murphy (1978) and Meadows (*in preparation*) found 'marks on' differences to be much smaller, up to half the amount, than 'marks off' differences, it is likely the study will reveal mark differences in the first sample much greater than 66 as used as a measure of acceptability in the live marking operation. The study explores the use of a mark difference threshold as an effective method for judging acceptability.

Whilst the change in procedure for the first phase sample may result in higher mark differences, it was not thought to be problematic. Meadows (*in preparation*) found that senior examiners used mark differences from both 'marks on' and 'marks off' scripts when deciding what feedback to give examiners, indicating that they take both seriously as indicators of marking quality. Baird, Greatorex and Bell (2003) showed that examiners not using 'marks off' exemplar scripts as part of the standardisation process were no less accurate than examiners who used them. Thus examiners who marked 'marks off' exemplar scripts and were sent them back with commentaries and feedback marked no less accurately than markers who were not required to mark them at all. Their findings support the assertion made by Shaw (2002) that the mark scheme alone, even without a standardisation meeting, exerts some standardising effect. This finding, however, was for examiners who were experienced and it could be an effect that is pertinent to them, but not to inexperienced examiners. Any differences in marking between experienced and inexperienced markers revealed in the present study will have to be interpreted with this in mind. Particularly as some studies have shown that inexperienced markers tended to mark more severely than experienced markers. Weigle (1999) found that prior to training, inexperienced markers could be significantly more severe than experienced markers depending on the essay prompt, 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 the rater training'. Ruth and Murphy (1988) report on a study that revealed a tendency for more severe marks from trainee teachers compared to those from experienced markers, though the differences were not significant. They suggest the markers' background determined the "distinctly different frames of reference for judging the essays". This seems a reasonable assertion and any differences in severity between the groups of markers in the study will be scrutinised closely.

A different aspect of marking accuracy is the rate at which markers make errors in the administration of marking. Inaccurate marking can be the result of the inappropriate application of the mark scheme, but equally it can be the result of an administrative error

made by the marker on the documentation. In Key Stage 3 marking, the onus to check marksheets and the answer booklets is on the marker before returning scripts to schools. A copy of the marksheet is sent to the External Marking Agency who checks that the marksheet columns have been completed before sending them to the Data Capture Agency. The latter agency inputs the marksheets into a database, verifies the accuracy of the conversion of marks to levels and sends the data to the DfES for the Performance Tables. Schools can apply to the External Marking Agency for a review of marks if they discover inaccuracies that would lead to a change in the final level of a pupil. Three types of review can be applied for:

- R1 clerical check on script and/or marksheet for individual pupils;
- R2 check of application of mark scheme for individual pupils;
- GR group review of scripts where there are substantial inaccuracies in the quality of marking that would impact on a school cohort.

An interest in maintaining clerical accuracy stems then not only from the intrinsic desire to see pupils awarded the marks they deserve, but also from the additional burden of responding to enquiries upon results. The process, though a necessary function of marking, is lengthy and costly.

The errors that are discussed here have high impact, in that they lead to a change in the final level a pupil receives. They carry importance from the point of view of the school because a pupil's correct level is at risk. From the External Marking Agency's point of view, they carry high significance because the accuracy of the markers needs to be assured, or corrected if it is errant. The current study only reports on clerical errors, but they are not the only errors that lead to a final level change, but any error. As a means of investigating the types of errors different types of markers may make, the clerical errors tallied were more in tune with AQA's mainstream examination checks (Pinot de Moira and Davies, 2002). The study reports the level of accuracy that different types of markers achieve in tallying marks and completing marksheets, with a view to gauging whether markers with different backgrounds might need different support or training.

By way of a summary, the main areas of interest in this study are marking reliability characterised by correlation measures between markers and the Lead Chief Marker and percentage of agreements in level assignation; marking accuracy as defined by the size of the difference in marks between a principal marker and a marker under study; and clerical accuracy. In order to contextualise the findings, the marker's attitudes towards marking were also collected and are reported.

2. METHODOLOGY

2.1 Participants

a. Lead Chief Marker

The Lead Chief Marker of Key Stage 3 English acted as the national standard in the study. He marked all of the scripts so that his marks could be compared to those of all other markers. In correspondence, he has pointed out when the marks for scripts used for training and standardisation are decided, it is done collectively between him and the chief and deputy chief markers. However, in the study, his single judgement was used as the best possible estimate for the pupils' true mark.

b. Markers

Four types of markers with an academic background in English but different amounts of teaching experience were identified:

1. BA English graduates of 2003,
2. PGCE Secondary English graduates of 2003,
3. Secondary English teachers with three or more years of teaching experience,
4. experienced Key Stage 3 markers who had not marked in 2003.

BA English graduates were recruited because they have the same subject background as Key Stage 3 English markers, but no experience of teaching or marking at all. BA graduates of English who had some teaching or marking experience could not apply. PGCE graduates were recruited because they had the same subject background and being about to complete their PGCE, they had a small amount of teaching experience gained on teaching practice. They represented markers with some teaching experience. Teachers of three or more years' experience were recruited because they were both familiar with the subject and had the requisite amount of teaching experience ideally sought in markers. Teachers with external marking experience could not apply so that marking experience would not be a confounding factor. They represented newly recruited markers with no external marking experience. Experienced markers were selected so long as they had been acceptable Key Stage 3 English markers before, but for whatever reason had not been available for live marking in 2003. They represented experienced markers coming to a marking session with no prior knowledge of the new marking scheme. They were the control group for some comparisons. Another group was considered, newly qualified teachers, but an operational study took place in 2003 to investigate their suitability as markers.

c. Team supervisors

Seven experienced markers from 2003 were recruited to the study 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. It was not possible to recruit team leaders for all of the supervisory roles because they were still involved in the live marking review operation and could not be distracted from their duties.

d. Trainers

Two trainers were recruited to run the two training days for the markers. They were the Lead Chief Marker and the Chief Marker for the southern region of the Key Stage External Marking Agency.

e. Pupils

The scripts of 100 pupils were selected at random from a population of 36,810 unmarked scripts held at the southern region External Marking. The sample was stratified to reflect the school types in the population. The number of pupils from each type of establishment is given in Table 2. The scripts were anonymised.

Table 2: Pupils in sample

Type of Establishment	% of pupils in population	Number of pupils in sample
1 Community school	63.0	65
2 Voluntary aided school	13.6	13
3 Voluntary controlled school	3.1	4
5 Foundation school	15.7	16
6 City technology school	0.5	1
7 Community special school	0.6	1
11 Other independent school	1.7	0
26 Overseas school	0.6	0
54 Offshore and overseas	0.5	0
Total	99.3*	100

* The remaining 0.7% of pupils was from types of establishment with very small entries.

Once the scripts had been marked, it was discovered that two of the scripts had provisional marks already written on them. Some markers ignored the marks and marked the scripts themselves, but other markers did not mark the scripts at all. Because of the difference in approach to these two scripts, the marks from all markers were removed from the data set making the eventual total number of pupils in the study 98. It is not possible to trace the type of establishment they were from.

2.2 Test materials

The test papers used in the study were the 2003 Key Stage 3 English papers. The markers were given photocopies of the pupils' entire work. The tests are delivered to the pupils as three **components**: reading, writing and Shakespeare. The components are made up of a number of **tasks**. The composition and maximum mark for each of the tasks in the three components are given in Table 3 below:

Table 3: Composition of Key Stage 3 English components in 2003

Component	No. of tasks	Number of tasks and sub-tasks out of *x* marks	Max. marks
Reading	13	8 tasks out of 1 mark 4 tasks out of 2 marks 2 tasks out of 3 marks 2 tasks out of 5 marks	32
Writing	1	1 task out of 30 marks	30
Shakespeare	2	1 reading task out of 18 marks 1 writing task out of 20 marks	38

This shows the configuration of the test that pupils see. The marks from the three components, however, are combined to give two **paper** scores to the pupil: a reading paper score and a writing paper score as shown in Table 4. The paper scores are further translated into levels according to the 2003 Key Stage 3 English Level Thresholds (given in Appendix 1), which are reported to the pupils.

Table 4: The components and tasks that constitute the reading and writing papers

Paper	Composition of paper	Max. marks
Reading	Reading component + Shakespeare reading task	50
Writing	Writing component + Shakespeare writing task	50

The two paper scores are further combined to give each pupil a **test** score out of 100 marks. There is also a level associated to the test score, which is also reported to the pupils.

There is, however, one more configuration of marks to consider. The markers are trained to mark the two writing tasks according to the mark scheme which specifies **strands**. The strands are amalgamations of assessment focuses described in the mark scheme. The strands and the number of marks allocated to each one are given in Table 5. In live marking, pupils see their marks at strand level on the front cover of the writing and Shakespeare components when they receive their scripts after the marking period. Throughout this report, the above definitions of task, component, paper, test and Key Stage level are used.

Table 5: Marking strands used to mark the writing tasks

Task	Strands	Max. marks
Writing component task	Sentence structure and punctuation (SSP)	8
	Text structure and organisation (TSO)	8
	Composition and effect (CEL)	14
Shakespeare writing task	Sentence structure, punctuation and text organisation (SSPTO)	6
	Composition and effect (CES)	10
	Spelling (S)	4

2.3 Recruitment of markers, team supervisors and trainers

The target number of markers per group was 20. The final number recruited is given in Table 6. They were recruited using a variety of methods. Advertisements for the first three types were placed in the national press and on the Internet in the first and second weeks of June 2003. The wording used is given in Figure 1. The English departments and PGCE English departments of eight universities and colleges were solicited directly to engender interest amongst graduating BA and PGCE students for the study prior to the break-up for the summer vacation. These efforts resulted in 63 formal applications across the first three types, at least 20 in each group. The application process mirrored that of the live marking operation as closely as possible. Thus the markers had to complete an application form and provide a reference from a professional who could vouch for their suitability to fulfil the requirements of a marker. Once the closing date for application was reached, 20 markers from each group were sent offer letters and the additional three markers were asked if they would wait in reserve in case the target number was not reached. The offer letters contained terms and conditions that followed those offered to live markers. Even making offers to the reserve markers did not yield the full complement of acceptances: 19 BA graduates accepted, 19 PGCE graduates accepted and 15 teachers accepted. And sadly, after offers were accepted,

a further five markers dropped out, two BA graduates and three PGCE graduates. This left a total of 17 BA graduates, 16 PGCE graduates and 15 teachers.

The recruitment of the experienced markers was approached in a different way. All markers on AQA's records as having marked for Key Stage 3 English before but who had decided not to mark in 2003 were approached directly by telephone. They proved to be the most difficult to recruit to the study because, firstly, it is a small pool of people to select from, and secondly, if markers became available for marking in 2003, live marking took precedence and they were offered an allocation of live scripts before they were offered a role in the study. All markers interested in participating in the study were sent offer packs that again mirrored almost exactly those used in live marking. Initially, 11 markers accepted the terms of the study, but sadly two had to drop out after accepting, resulting in nine experienced markers.

Table 6: Number of participants recruited to the study

Marker groups	Number
BA English graduates	17
PGCE Secondary English graduates	16
English teachers	15
Experienced Key Stage 3 English markers	9

The team supervisors and trainers were approached directly by telephone and were similarly required to accept the terms described in an offer letter.

Figure 1: Advertisements to recruit markers

Wanted

BA (English) graduates of 2003

PGCE (Secondary English) graduates of 2003

or

Secondary English teachers with three years of teaching experience but no external marking experience

For a research study

- Are you one of the above?
- Are you prepared to attend training meetings on 19th July and 26th July 2003?
- Are you willing to mark 100 exam papers at home by 20th August 2003?
- Are you interested in earning up to £465 for the training and marking plus expenses?

If you can answer yes to all of the above, please apply for further details by 13th June 2003

2.4 Training and training materials

Prior to the first training day, all markers were issued materials and training scripts relevant to preparation for marking. The materials used were the same as those that had been used in the live operational marking with the 'Administration File for markers and supervisors taking part in the Key Stage 3 English marker study' adapted to the timings and requirements of the study.

Two training days were held. The first one, on 19th July 2003, covered the reading paper, the Shakespeare reading task and marking administration. The second training day, on 26th July 2003, covered the writing paper and the Shakespeare writing task. The training methods were exactly those used in the live training days. The team supervisors acted as table managers to the markers allocated to their team, between eight and nine markers per team.

The markers followed up both training days by completing the exercises in the training pack. These exercises consisted of copied pieces of work from pupils for the markers to mark and send to their team supervisor. They were the same pieces of work used in the live operational marking. The team supervisors checked the marking and sent feedback to the markers on their marking using the standardisation procedures for live marking.

2.5 Marking

The markers each received the photocopies of complete Key Stage 3 English scripts of the 100 pupils at their homes. They marked at their own pace after the second training day and were given the deadline for completion of 22nd August 2003.

2.6 Standardisation

The markers were required to send the marked scripts and marksheet of pupils numbered 45 to 66 to AQA as a first sample check. The team supervisors were not used for this procedure because as experienced markers they were busy with marking review duties in the live marking operation. Instead, the markers received feedback in the form of commentaries on ten of the 22 scripts which had been prepared by the Lead Chief Marker. The commentaries included his marks and his reasoning for the mark he allocated. This procedure deviated from that used in the live operational marking. Normally, the markers would send their first sample to their team leader who would mark 10 of the 22 scripts and provide feedback. In the study, the markers all had the same scripts which enabled the same ten scripts to be used as the first sample. Furthermore, the lack of availability of the team supervisors for this procedure required an alternative method of standardisation.

2.7 Keying the marks

Once all of the scripts and marksheets had been returned to AQA, the marksheets were separated from the scripts, copied and the originals were sent to an external data keying agency. The agency keyed in the component and task level marks only, not the totals and levels, which could be calculated accurately later. This formed the uncleaned marks dataset.

2.8 Error checking

Two sources of errors were investigated: the scripts and the marksheets. Until summer 2003, AQA conducted checks of all scripts in the mainstream examinations (ELC, GNVQ, GCSE,

GCE and VCE). The procedure changed in the summer of 2003 to only a sample of scripts being checked. Since Key Stage 3 English procedures do not include any element of script checking at a central location, the procedures for checking scripts in the study were borrowed from the mainstream operation. Marksheets are not used in any of the mainstream examinations, and, again, there is no precedent for checking them centrally in Key Stage 3 English procedures, so a new procedure was developed for the study.

The types of errors checked were:
Script errors

I. Not all the work marked
II. Within script: question total missing - reading component only
III. Within script: addition error in question total - reading component only
IV. On front mark-box: question total missing
V. On front mark-box: wrong question total carried over from script - reading component only
VI. On front mark-box: column/component total added up wrongly

Marksheet errors

A. Total written when marks incomplete
B. Total blank when all marks present
C. Total incorrect
D. Level written when marks incomplete
E. Level blank when all marks present
F. Level incorrect
G. Marks missing

There were two script checks: the first check was for any errors made on the scripts and the second check was to ensure that the correct marks had been keyed into the dataset for analysis. All scripts were checked for errors. There were 100 pupils allocated to each of the 58 markers and there were three components per pupil, reading, writing and Shakespeare. This makes a total of 17,400 scripts. The scripts were bundled according to the pupils on the marksheets which ran to four bundles of 22 scripts and one bundle of 12 scripts per marker. In the first check, temporary clerical staff checked one bundle at a time, working through every script recording the errors listed above. They recorded every error they found on the Script Error Checking Form in Appendix 2A. If they found a transfer or addition error, they wrote in the correct mark or total in pencil on the script as well as recording it on the form. The second check was done against a print-out of the marks. Since the scripts had been corrected, any differences between the script and keyed marks were interpreted as errors in the keyed data. The print-out was corrected if an error was found. The dataset was then corrected accordingly. The errors recorded on the forms were keyed in.

All marksheets were checked. A print-out of totals and levels calculated using the uncleaned keyed data was used to check photocopies of the marksheets using the list of marksheet errors above. It was assumed that the uncleaned data was an error-free record of the marksheets, as assurance was given by the data keying agency. Any errors found were written on the copied marksheet and recorded on the Marksheet Error Form in Appendix 2B.

There was no check of the marksheets against the scripts, which would have provided an estimate of transfer errors, because it had been noted at the project initiation stage that error checking was of less importance than other aspects of the project. The errors recorded on the forms were keyed in.

2.9 Qualitative data collection

Near to the end of the marking period, each marker was asked to complete a questionnaire (see Appendix 3) that asked for their opinions on all aspects of the marking process, from start to finish. The responses in the returned questionnaires were coded.

3. RESULTS

3.1 General comparison of the raw marks

Each marker marked the same set of 98 scripts. If there was no difference in marking between any of the markers, the same marks would be obtained across the allocation of scripts regardless of marker type. As a summary measure, the mean mark for the entire allocation was calculated for each marker type and is summarised in Table 7. The presentation of the markers' means as group means masks the variation within the groups, that is, variation within the between-subject groups. Figures 2a-g contains seven plots indicating the means per marker. The mean of the marker type means is also shown. The means were plotted as a percentage of the maximum mark possible on the test, papers, components and tasks so that the graphs can be compared.

The plots in Figures 2a-g indicate variability within the marker groups and tests of homogeneity of variance were carried out to see if there were differences in variability between the groups. The tests used the *F*-test for the ratio of the larger variance of a pair to the smaller (Howell, 1992). The results are given in full in Appendix 4, but to summarise, homogeneity of variance was found in the majority of comparisons, including all of the reading component and the Shakespeare reading task comparisons. This indicates that the raw marks from the different types of marker in the assessments were from the same population. Instances of heterogeneity of variance were found in only four comparisons, each one involving the BA graduates and either the teachers or the experienced markers. In these, the BA graduates yielded a distribution of marks that was significantly larger than that of one other group of markers.

To test for differences in the leniency (or severity) between the groups' raw marks, repeated measures analyses of variance were conducted. The pupils' raw scores were the 98 within-subject variables and the four groups constituted 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 own independent error terms. For the purposes of the study, the within-subject analysis, including the interaction term, 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 did not violate the assumptions. The results of all the analyses are given in Appendix 4, but to summarise, the tests indicated virtually no between-subject effects. The only significant differences detected by the tests were in the composite scores of the reading paper ($F_{3,38}$=2.954, p=0.045).

Indeed, the groups' means in Table 7 indicate that the experienced markers were more lenient than the other groups. No other aspect of the test indicated significant differences between the marker groups, suggesting there were no differences in the lenience (or severity) of the marker groups.

Table 7: Mean marks across the entire allocation by marker type

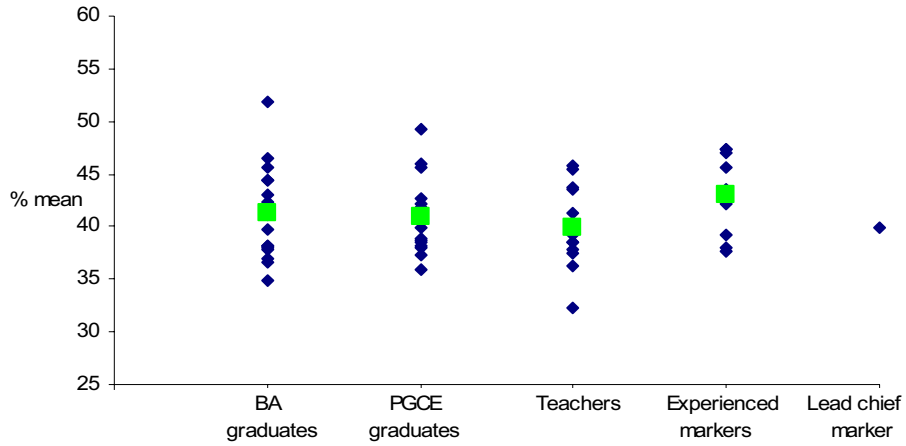
		N	Mean	Standard deviation
English test total (100 marks)	BA graduates	1646	41.23	17.14
	PGCE graduates	1520	40.82	17.72
	Teachers	1462	39.92	18.11
	Experienced markers	870	43.17	18.02
	Lead Chief Marker	*97*	*39.81*	*16.84*
Reading paper (50 marks)	BA graduates	1648	21.66	8.99
	PGCE graduates	1521	21.39	9.33
	Teachers	1464	21.39	9.38
	Experienced markers	870	22.52	9.56
	Lead Chief Marker	*97*	*20.01*	*8.81*
Writing paper (50 marks)	BA graduates	1661	19.49	9.27
	PGCE graduates	1565	19.44	9.48
	Teachers	1467	18.49	9.80
	Experienced markers	876	20.58	9.43
	Lead Chief Marker	*98*	*19.70*	*9.23*
Reading component (32 marks)	BA graduates	1655	14.67	5.98
	PGCE graduates	1549	14.62	6.25
	Teachers	1466	14.60	6.05
	Experienced markers	870	15.16	6.30
	Lead Chief Marker	*97*	*13.41*	*5.93*
Writing component (30 marks)	BA graduates	1664	10.82	5.73
	PGCE graduates	1567	10.78	5.87
	Teachers	1468	10.13	5.99
	Experienced markers	876	11.40	5.75
	Lead Chief Marker	*98*	*10.39*	*5.06*
Shakespeare reading task (18 marks)	BA graduates	1659	6.94	3.82
	PGCE graduates	1540	6.76	3.79
	Teachers	1468	6.78	3.96
	Experienced markers	876	7.34	3.92
	Lead Chief Marker	*98*	*6.57*	*3.48*
Shakespeare writing task (20 marks)	BA graduates	1663	8.65	4.46
	PGCE graduates	1566	8.65	4.53
	Teachers	1469	8.36	4.70
	Experienced markers	876	9.18	4.56
	Lead Chief Marker	*98*	*9.32*	*4.92*

In order to locate the source of the difference in the reading paper, *a priori* two-sample t-tests were carried out on three pairings: the experienced markers against each of the other three groups. The results revealed that the experienced markers were more lenient than the PGCE graduates and teachers (t=-2.03, df=2417, p=0.04 and (t=-2.13, df=2334, p=0.03, respectively). There was no significant difference between the BA graduates and experienced markers (t=-1.92, df=2523, p=0.06). These findings suggest that having teaching but no marking experience may have contributed to markers being more severe than experienced markers on this paper, but having no experience of either teaching or marking contributed to markers being similarly lenient as experienced markers. It should be stressed that, on all other aspects of the test, no significant differences in lenience were found, suggesting that for the majority of the assessment, having teaching experience or none at all made no difference to the lenience of a marker. These findings are only of marginal interest because they measure the lenience of the groups relative to each other and not against a standard. It is noticeable in Table 7 that Lead Chief Marker is more severe than the average of other markers in each composition of the test, the components and tasks, except in the Shakespeare writing task, where he was the most lenient. Thus the size of the deviation of the markers' marks from those of the Lead Chief Marker give us a measure of accuracy for each group as measured against a standard. This is investigated in more detail in the section investigating mark differences in Section 3.5.

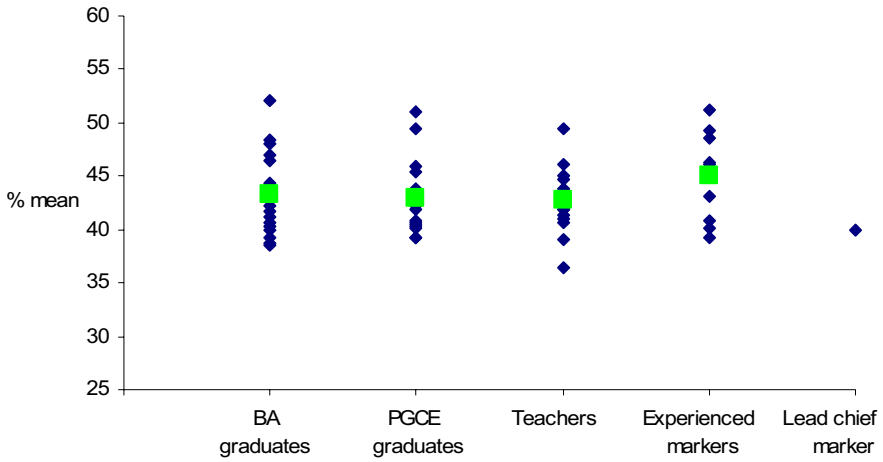
Figures 2a-g: Mean mark on the entire allocation per marker for the test, components, tasks and papers

Legend
♦ = mean mark of individual markers
■ = mean of means per marker type

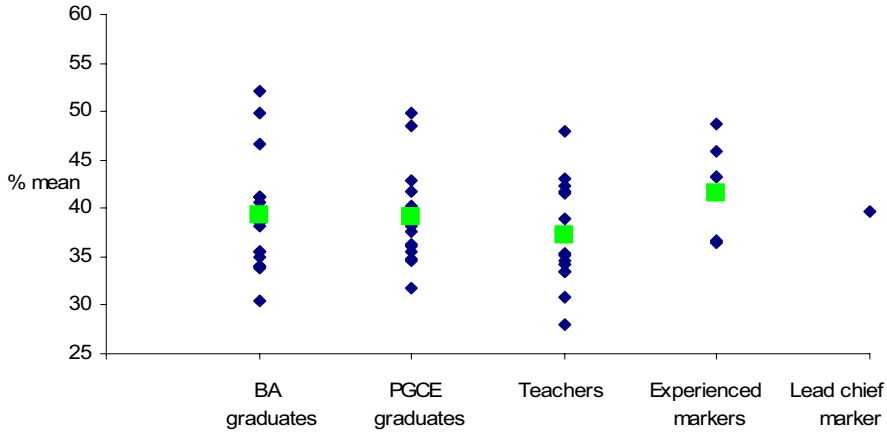
2a. English test total (max mark = 100)



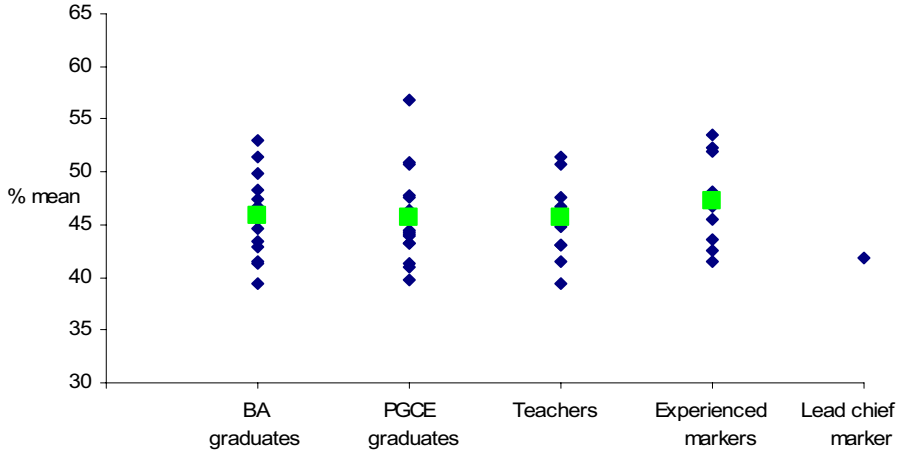
2b. Reading paper mean marks (max mark = 50)



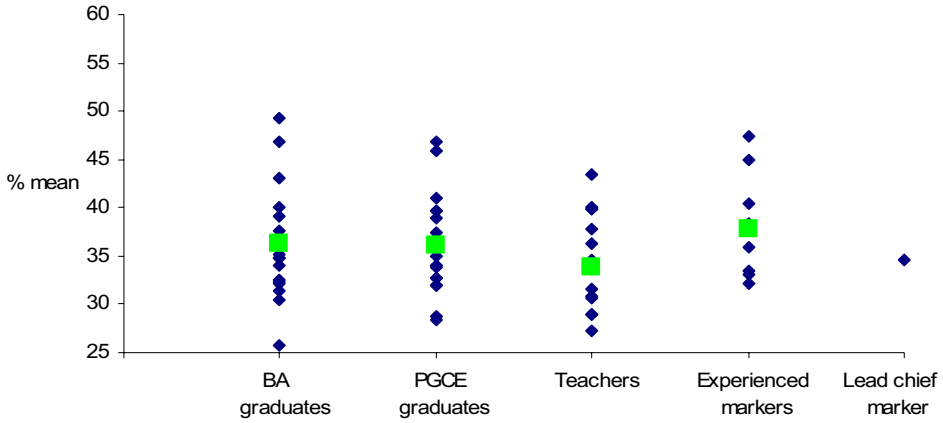
2c. Writing paper mean marks (max mark = 50)



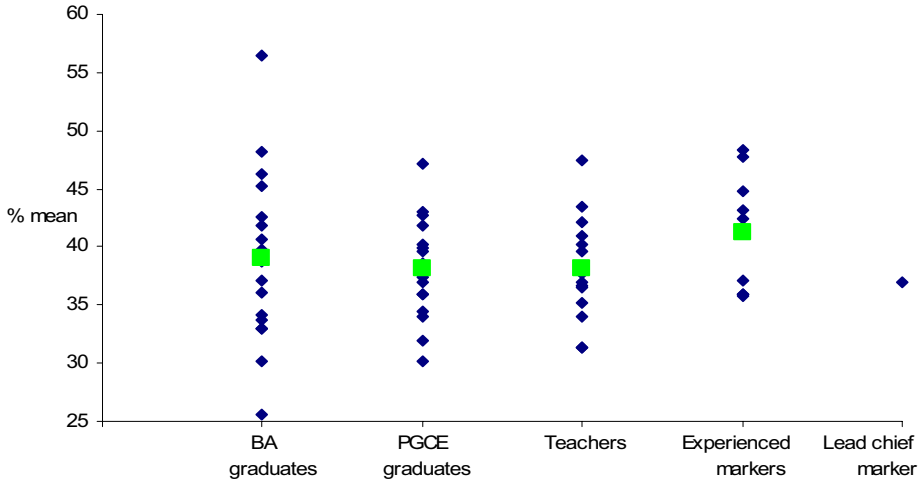
2d. Reading component mean marks (max mark = 32)



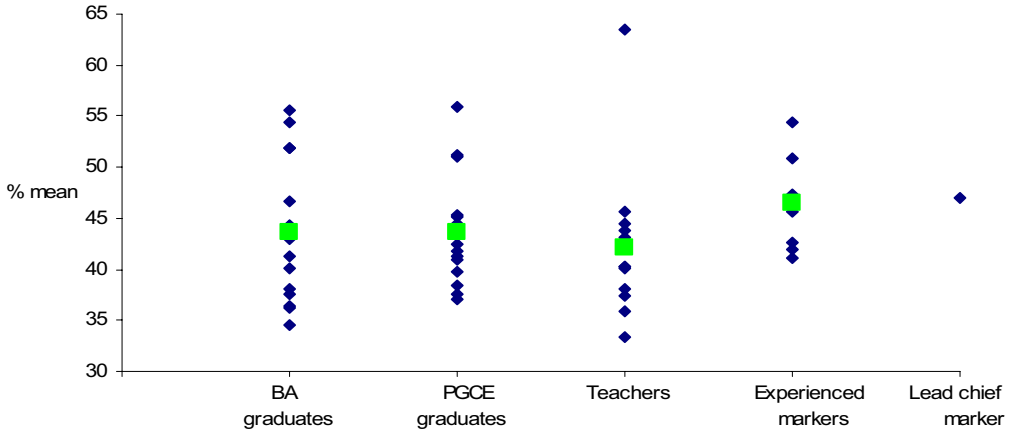
2e. Writing component mean marks (max mark = 30)



2f. Shakespeare reading component mean marks (max mark = 18)



2g. Shakespeare writing component mean marks (max mark = 20)



3.2 Reliability estimates

Multiple markings of the same scripts can lead to dozens of inter-rater comparisons. Since the Lead Chief Marker's marks represent the national standard, of interest here are the markers versus Lead Chief Marker comparisons. Pearson product-moment correlation coefficients were calculated on the overall test scores, the two paper scores and the two component and two tasks scores. Table 8 summarises the comparisons by showing the mean of the correlation coefficients by group, the standard deviations, and the highest and lowest coefficients in the group. To get a better idea of how the reliability estimates varied within and between groups, the coefficients were plotted in Figures 3a-g.

Table 8: Summary of correlation coefficients between each marker and Lead Chief Marker

	N	Mean	SD	Minimum	Maximum
English test					
BA graduates	17	0.89	0.03	0.85	0.95
PGCE graduates	16	0.91	0.02	0.87	0.94
Teachers	15	0.92	0.02	0.89	0.96
Experienced markers	9	0.92	0.01	0.90	0.94
Reading paper					
BA graduates	17	0.92	0.02	0.88	0.97
PGCE graduates	16	0.93	0.01	0.91	0.95
Teachers	15	0.94	0.01	0.91	0.96
Experienced markers	9	0.94	0.01	0.93	0.96
Writing paper					
BA graduates	17	0.75	0.07	0.63	0.86
PGCE graduates	16	0.77	0.04	0.71	0.82
Teachers	15	0.80	0.04	0.72	0.87
Experienced markers	9	0.80	0.04	0.74	0.85
Reading component					
BA graduates	17	0.91	0.02	0.88	0.95
PGCE graduates	16	0.91	0.02	0.87	0.93
Teachers	15	0.91	0.02	0.89	0.94
Experienced markers	9	0.91	0.01	0.88	0.93
Writing component					
BA graduates	17	0.63	0.10	0.40	0.77
PGCE graduates	16	0.65	0.05	0.51	0.72
Teachers	15	0.69	0.08	0.48	0.81
Experienced markers	9	0.67	0.07	0.52	0.76
Shakespeare reading task					
BA graduates	17	0.78	0.07	0.65	0.90
PGCE graduates	16	0.81	0.06	0.69	0.89
Teachers	15	0.85	0.03	0.77	0.89
Experienced markers	9	0.86	0.02	0.82	0.89

Shakespeare writing task					
BA graduates	17	0.74	0.07	0.53	0.84
PGCE graduates	16	0.77	0.05	0.66	0.84
Teachers	15	0.79	0.05	0.65	0.85
Experienced markers	9	0.80	0.03	0.75	0.85

Tests for the homogeneity of variance of the reliability estimates between the four groups were carried out using the *F*-test as described in Section 3.1. The results are given in full in Appendix 4, but to summarise, homogeneity of variance was found in the majority of the comparisons, including all comparisons of the test, the reading component and the Shakespeare writing task. This suggests the reliability estimates for these three sets of marks were from the same population. Figures 3a, 3d and 3g support the assertion that the variation between the groups is of a similar magnitude. Instances of heterogeneity of variance were found in five comparisons in other aspects of the test, all of them involving the BA graduates and either the PGCE graduates or the experienced markers. Thus, on five comparisons in other aspects of the test, the BA graduates yielded a distribution of reliability estimates that was significantly larger than that of one or two other groups of markers. Comparisons between the other marker groups suggested homogeneity of variance, that is, reliability estimates were distributed in a similarly large or small way.

In order 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[2] using the Welch procedure to take account of the unequal sample sizes and the instances of heterogeneity of variance. The full results are given in Appendix 4. No significant differences were found for the reading paper, the writing paper, nor the reading and writing components. This suggests the reliability estimates at paper and component level from the different groups of markers were indistinguishable. Significant results were found at test level and for the two Shakespeare tasks. In order to determine whether teaching experience was a contributing factor to the differences in reliability estimates, *a priori* t-tests were carried out between the experienced markers and the other three groups. The results are summarised in Table 9. At test level, both BA and PGCE graduates 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 indicate that at least three years of experience contributed to higher reliability estimates. Only the BA 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.

[2] 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).

Table 9: Results of *a priori* t-tests to test for differences in reliability estimates

	Comparing experienced markers against ….	*t*	df	*p*
English test	BA graduates	-2.56	24	0.02
	PGCE graduates	-2.30	23	0.03
	teachers	-0.10	22	0.92
Shakespeare reading task	BA graduates	-3.28	24	<0.01
	PGCE graduates	-2.23	23	0.04
	teachers	-1.06	22	0.30
Shakespeare writing task	BA graduates	-3.01	23.01*	0.01
	PGCE graduates	-1.62	23	0.12
	teachers	-0.74	22	0.46

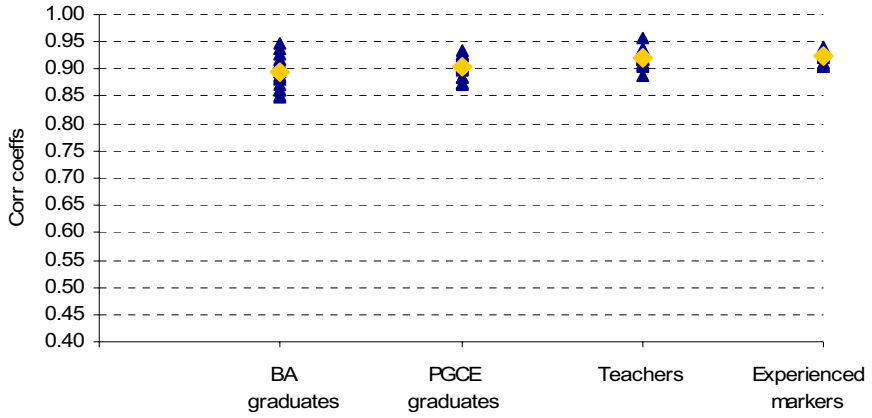
*degrees of freedom adjusted to take unequal variances into account

To summarise this section, 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 PGCE graduates, was found to be a contributing factor to significantly higher reliability estimates on the Shakespeare writing task. No other comparisons indicated that teaching experience or a lack thereof made any difference to the reliability estimates. These results suggest that for some reason the two Shakespeare tasks pose difficulties to markers who do not have teaching experience.

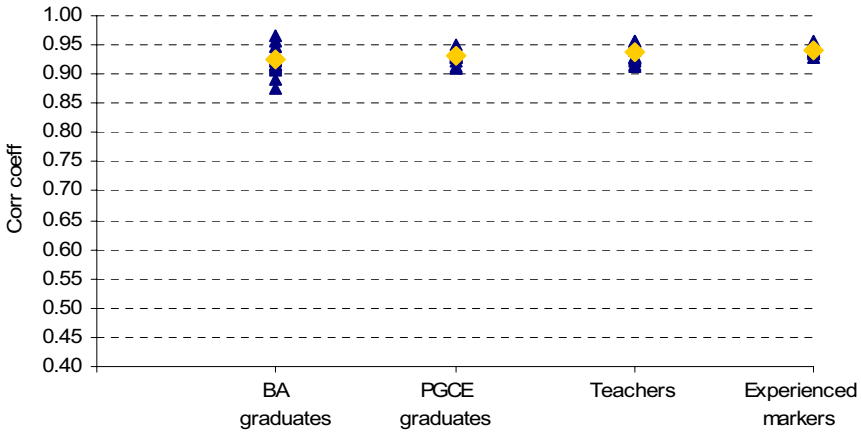
Figures 3a-g: Plots of the correlation coefficients of each marker against the Lead Chief Marker by group

Legend ▲ = correlation coefficient of a marker and the Lead Chief Marker
♦ = mean correlation per group

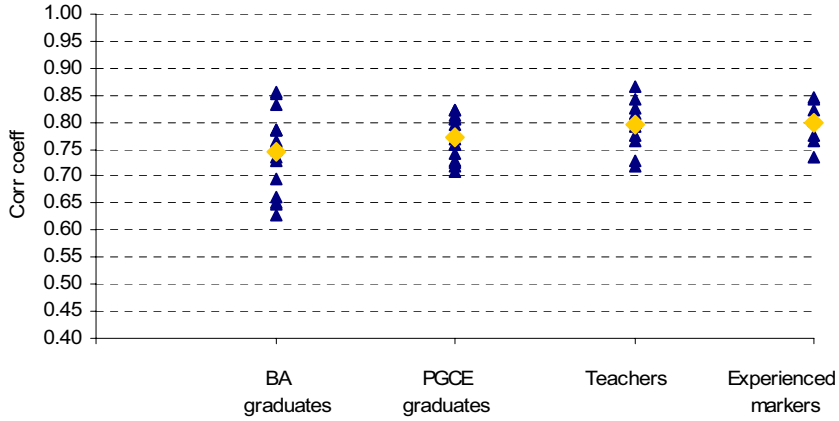
3a. Test scores



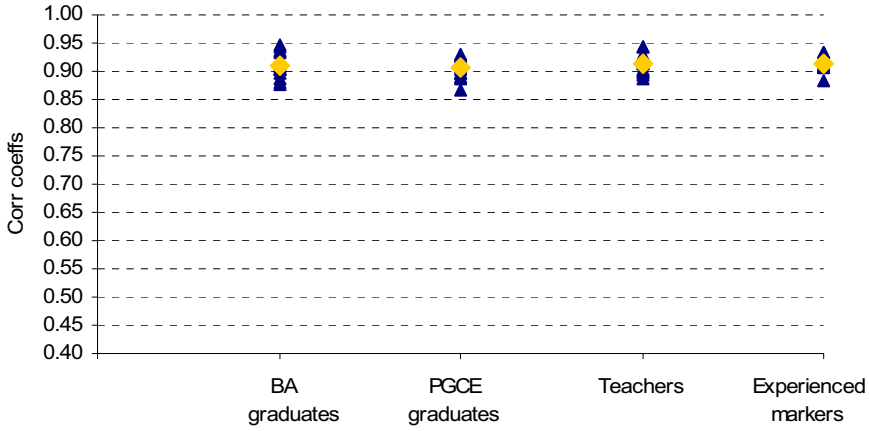
3b. Reading paper scores



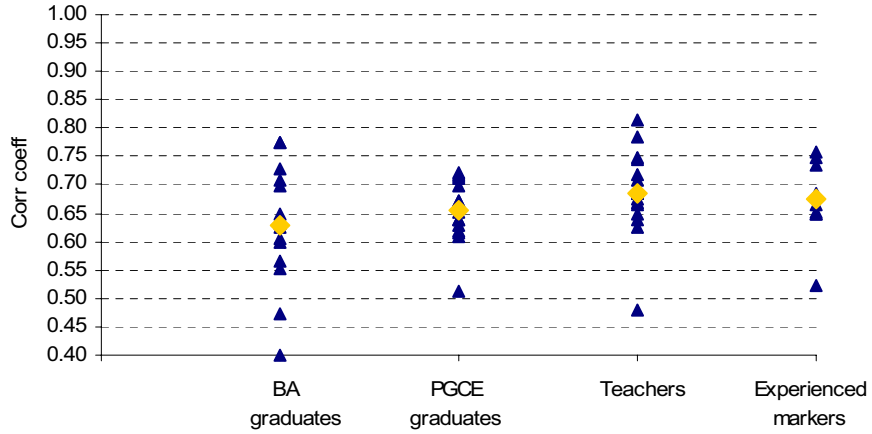
3c. Writing paper scores



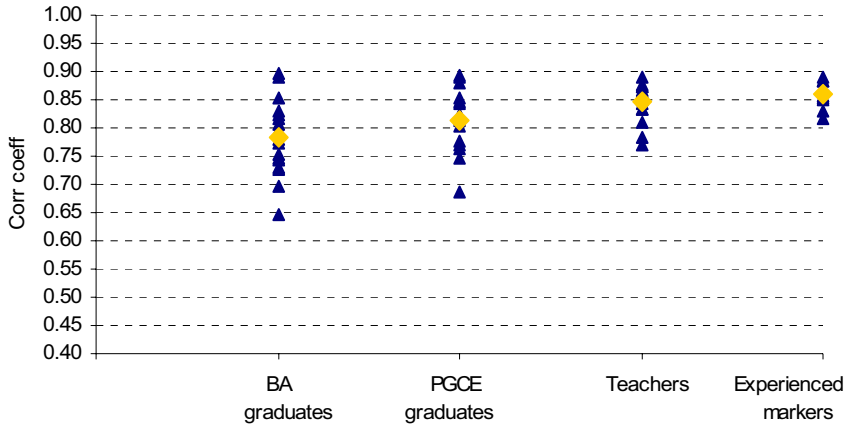
3d. Reading component scores



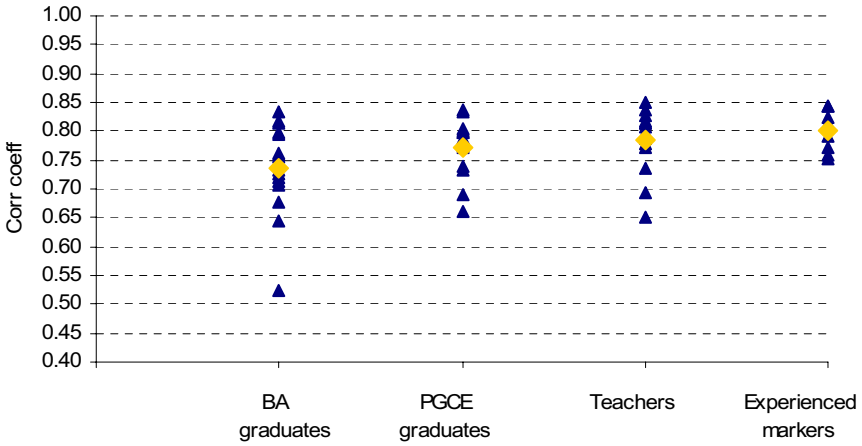
3e. Writing component scores



3f. Shakespeare reading task scores



3g. Shakespeare writing task scores



3.3 Percentage of same Key Stage levels awarded by markers and Lead Chief Marker

Another measure of the reliability of marking is the number of agreements in the Key Stage levels awarded to pupils by the markers and Lead Chief Marker. Three levels are assigned to each pupil: one for the reading paper, one for the writing paper and one for the test overall. Reading and writing scores are awarded at four levels: 4, 5, 6 and 7. The test scores are awarded at six levels: N, 3, 4, 5, 6, and 7. Key Stage levels for 2003 are given in Appendix 1. Table 10 shows the percentage of same levels assigned by the four groups for the test, reading paper and writing paper scores. For levels awarded to the test scores across all markers, the percentage of agreements was 61.22% and the percentage of agreements within one grade was 97.67%. For the reading levels, the rates were 65.30% and 98.22%; and for the writing levels, they were 50.22% and 94.16%.

To test whether the mean number of agreements was the same for the four marker groups, an analysis of variance was carried out on the number of agreements per marker for each of the three levels. No significant differences were found (test levels Welch statistic=0.811, df=3, 25.20, p=0.52; reading levels Welch statistic=0.75, df=3, 24.86, p=0.53; and writing levels Welch statistic=0.74, df=3, 26.31, p=0.54). These results suggest that there was no difference between the groups with regard to their accuracy at assigning pupils the same levels as the Lead Chief Marker.

Table 10: Percentage of same levels assigned by the four marker types

Marker type	Test	Reading	Writing
BA graduates	58.69	64.64	48.28
PGCE graduates	61.45	66.71	50.99
Teachers	62.65	67.22	50.37
Experienced markers	58.85	60.83	52.28
All markers	61.22	65.30	50.22

The direction of the misclassifications was also investigated to test whether the groups were similarly lenient or severe. Across all markers, the tendency was for leniency on the English test levels, with 21.39% of markers assigning one level or more higher than the Lead Chief Marker. Table 11 shows that the experienced markers tended to be more lenient than the Lead Chief Marker, as to be expected because their overall marks were more generous. The other three marker types appeared equally generous and severe. However, these differences were not significant (χ^2_{5498} = 57.60, p>0.05, see Appendix 4 for the raw counts for Table 10). Similar tests were run for the reading and writing levels, but again no significant results were found even though the tendency was also for lenience (reading χ^2_{5452} =35.87, p>0.05, writing χ^2_{5569} =64.21, p>0.05, raw counts in Appendix 4).

Even though the BA graduates were found to have lower reliability estimates than the teachers on the test scores and the two Shakespeare tasks, there was no difference with regard to the accuracy with which they assigned the pupils' levels. The pattern of leniency or severity is no different between the marker groups. The experienced markers do not appear to be any more accurate than the other types of marker compared to the levels assigned by the Lead Chief Marker on either paper or on the test overall.

Table 11: Proportion of same and adjacent levels awarded for the overall test

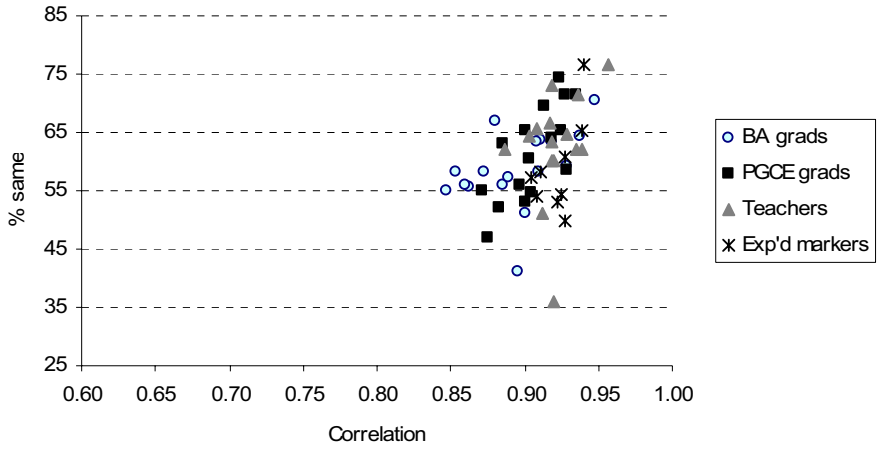
	2 levels lower	1 level lower	Same	1 level higher	2 levels higher	3 levels higher
BA graduates	1.40	17.07	58.69	21.93	0.85	0.06
PGCE graduates	1.51	16.18	61.45	19.93	0.92	0.00
Teachers	2.33	17.37	62.65	17.03	0.62	0.00
Experienced markers	0.69	12.07	58.85	27.59	0.80	0.00
Total	1.54	15.84	61.22	20.61	0.79	0.02

3.4 Relationship between reliability estimates and percentage same level

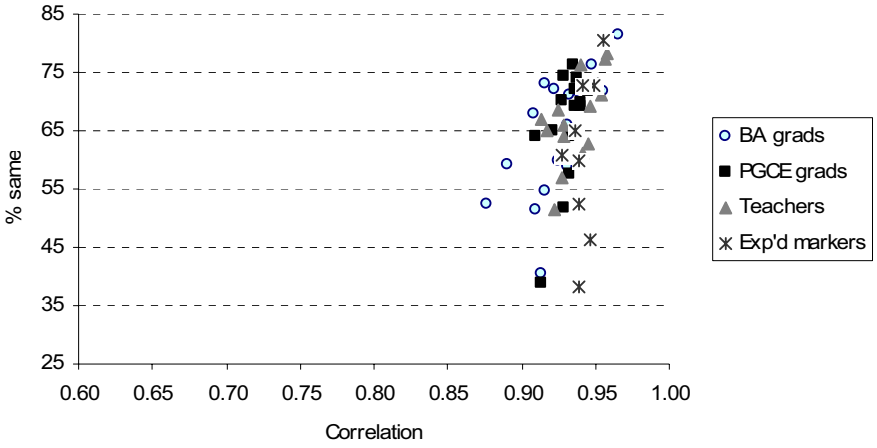
The results in the previous section suggest the relationship between the reliability estimates and the proportion of same levels awarded by the markers and Lead Chief Marker is not straightforward. Higher reliability estimates do not appear to result automatically in more accurate assignation of levels compared to lower reliability estimates. Figures 4a-c plot the reliability estimates against the proportion of levels awarded the same by each marker and the Lead Chief Marker for the test, reading and writing papers. There is a degree of positive correlation between the reliability and percentage same estimates (test scores, r=0.445, p<0.01; reading paper, r=0.494, p<0.01; and writing paper, r=0.540, p<0.01). A glance at the plots suggests that no one group stands out as having a particular profile different to any other group. The experienced markers tended to have more tightly grouped reliability estimates, but their percentage same estimates were still spread out, this is particularly so on the reading paper (Figure 4b) where theirs were amongst the lowest percentage same estimates. BA graduates yielded the lowest reliability estimates on the writing paper and they dominate the lower left hand corner of the plots (Figures 4a-c), which indicate both low reliability and percentage same estimates. Yet, other BA graduates feature in other parts of the plots.

Figures 4a-c: Reliability against proportion of same levels awarded by marker type

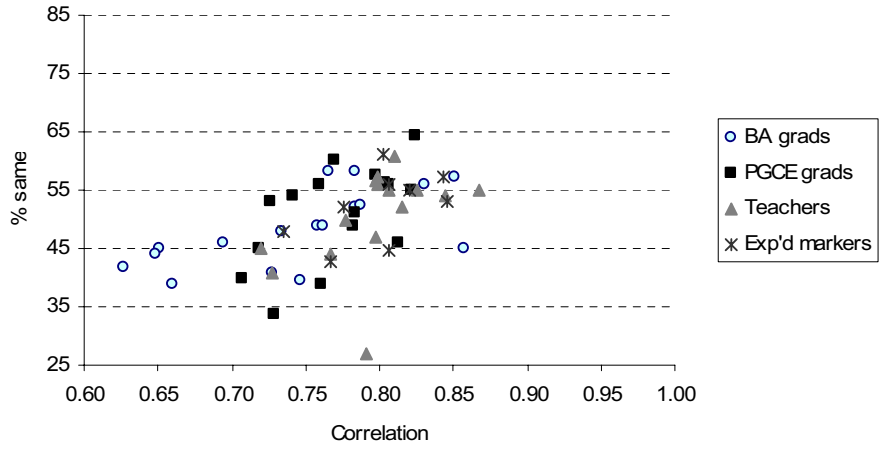
4a. Test scores



4b. Reading paper



4c. Writing paper



Whilst the relationship between reliability estimates and percentage same levels is positively correlated and significantly different from zero, higher reliability estimates are not necessarily associated with a higher degree of accuracy with respect to the number of same levels awarded by a marker to a pupil. The relationship is not clearly defined and further analysis is recommended. The accuracy of the level that is awarded to a pupil is the key indicator of the reliability of the test, irrespective of the marking reliability estimates demonstrated by different markers. In this study, there was no difference in the accuracy of levels assigned by the four groups compared to the national standard. However, only 61% of the levels were awarded accurately.

3.5 Marking accuracy

Marking accuracy was investigated in terms of the differences between a marker's marks and those of the Lead Chief Marker. Mark differences were calculated at two stages: the first was the first phase sample and the second was the entire allocation of scripts. The second comparison included alterations to marks based on the feedback from the first sample and

thus represented the markers' best estimate of overall accuracy as compared to the Lead Chief Marker. The first phase sample mark differences are investigated further in Section 3.6.

The dependent variable used to measure accuracy was the absolute mark difference, that is, the difference between a marker's mark and that of the Lead Chief Marker for every pupil recorded as a non-negative quantity. The mark difference, that is, the difference indicating whether the differences were positive or negative, would, in effect, be the same as the raw marks data decreased by the same constant for each pupil. The raw marks data have already been investigated in Section 3.1 and so the mark differences are not investigated further here.

Analysing the absolute mark differences will tell us whether there are any differences in accuracy between the groups because larger mark differences are an indication of a greater deviation away from the Lead Chief Marker, the national standard for the purposes of the study. Table 12 shows the mean absolute mark differences for the test scores, the paper scores, the two components and the two Shakespeare task scores by marker type. The differences in the size of the mean absolute mark difference per marker are shown clearly in Figures 5a to g.

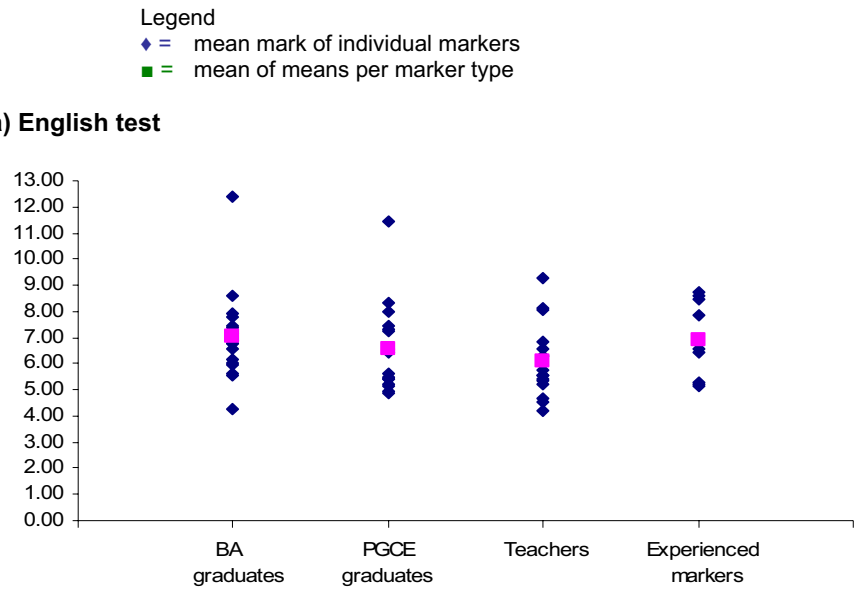
Table 12: Absolute mark differences by marker type

		N	Mean	Standard deviation
English test total (100 marks)	BA graduates	1630	7.01	5.59
	PGCE graduates	1507	6.48	5.30
	Teachers	1447	6.09	5.07
	Experienced markers	863	6.93	5.18
Reading paper (50 marks)	BA graduates	1632	3.27	2.79
	PGCE graduates	1508	3.05	2.56
	Teachers	1449	2.97	2.51
	Experienced markers	863	3.61	2.85
Writing paper (50 marks)	BA graduates	1695	5.54	4.33
	PGCE graduates	1597	5.23	4.28
	Teachers	1497	5.17	4.27
	Experienced markers	894	4.96	3.87
Reading component (32 marks)	BA graduates	1639	2.33	1.94
	PGCE graduates	1536	2.43	2.03
	Teachers	1451	2.22	1.90
	Experienced markers	863	2.63	2.20
Writing component (30 marks)	BA graduates	1698	3.76	3.10
	PGCE graduates	1599	3.76	3.10
	Teachers	1498	3.62	2.99
	Experienced markers	894	3.59	3.00

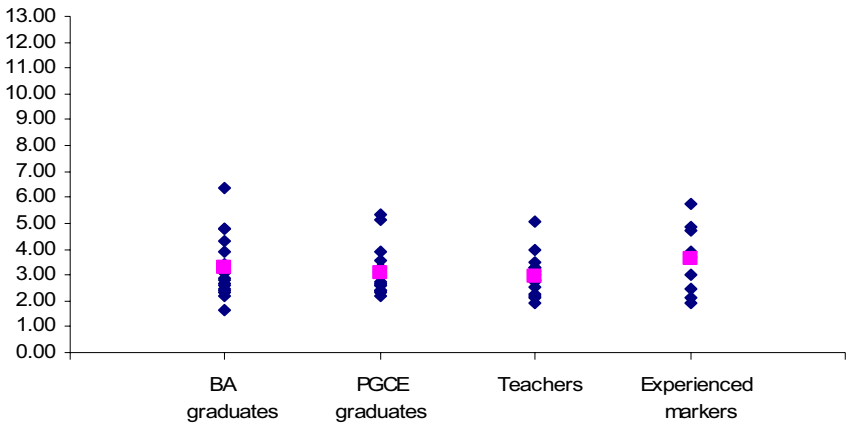
Shakespeare reading Task (18 marks)	BA graduates	1693	2.08	1.72
	PGCE graduates	1570	1.83	1.49
	Teachers	1498	1.73	1.47
	Experienced markers	894	1.72	1.55
Shakespeare writing Task (20 marks)	BA graduates	1697	2.89	2.25
	PGCE graduates	1598	2.73	2.06
	Teachers	1499	2.73	2.25
	Experienced markers	894	2.46	1.91

Tests for homogeneity of variance (see Appendix 4 for the full results) indicated that the absolute mark differences were similarly distributed between the different marker types only on the writing component (Figure 5e). Instances of heterogeneity of variance were found in all other aspects of the test for most, not all, of the comparisons. The pattern of differences in variability is complex and does not suggest one group is consistently more variable than the others on all aspects of the test. To summarise, for some reason, the accuracy of the BA graduates is more variable on the test overall and on the Shakespeare reading component. The accuracy of the experienced markers is more variable on the reading component, but less variable than any of the other groups on the writing paper and the Shakespeare writing task.

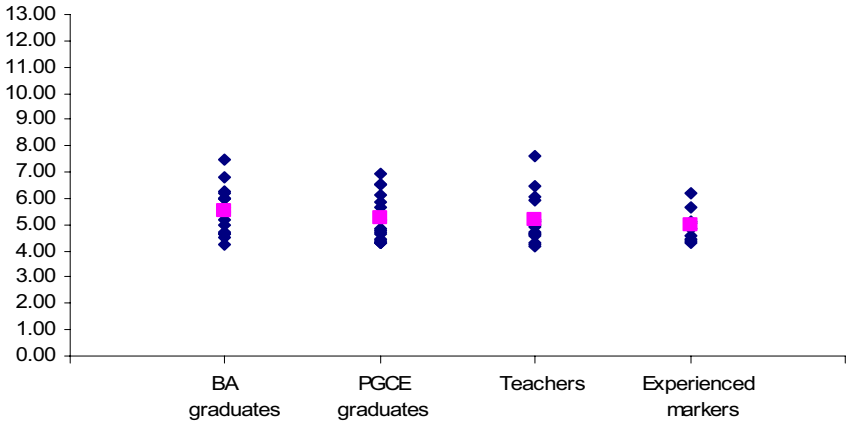
Figures 5a-g Plots of mean absolute mark difference per marker



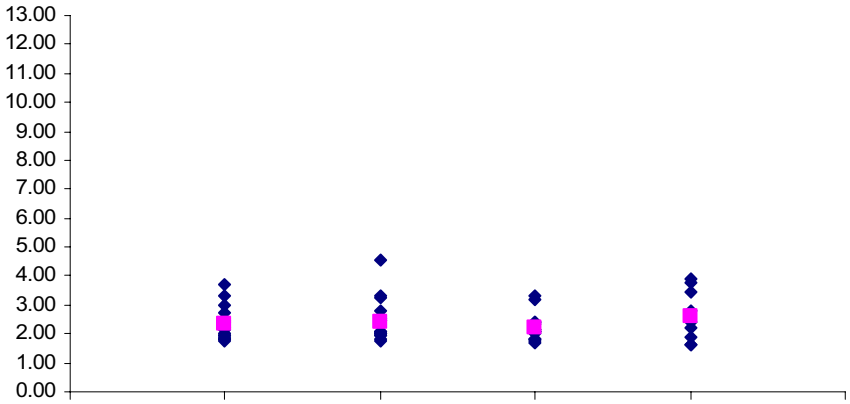
(b) Reading paper



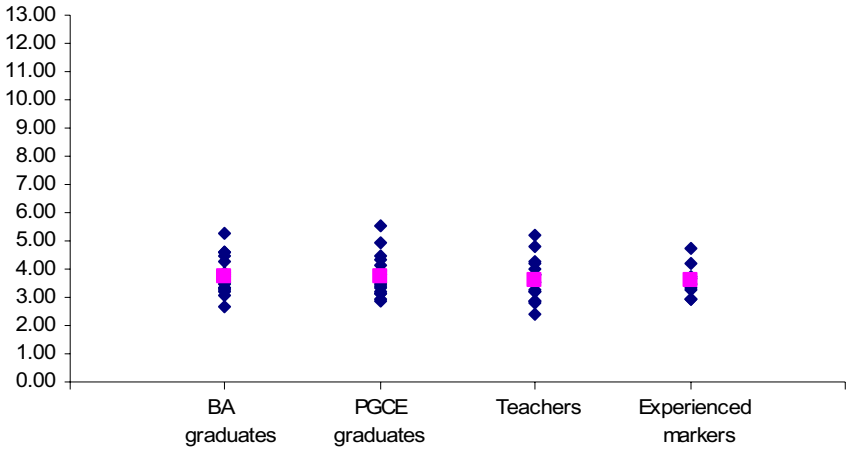
(c) Writing paper



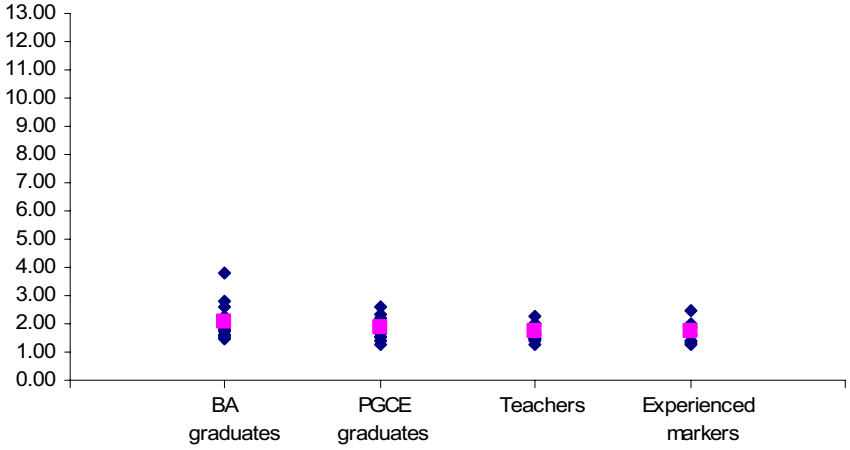
(d) Reading component



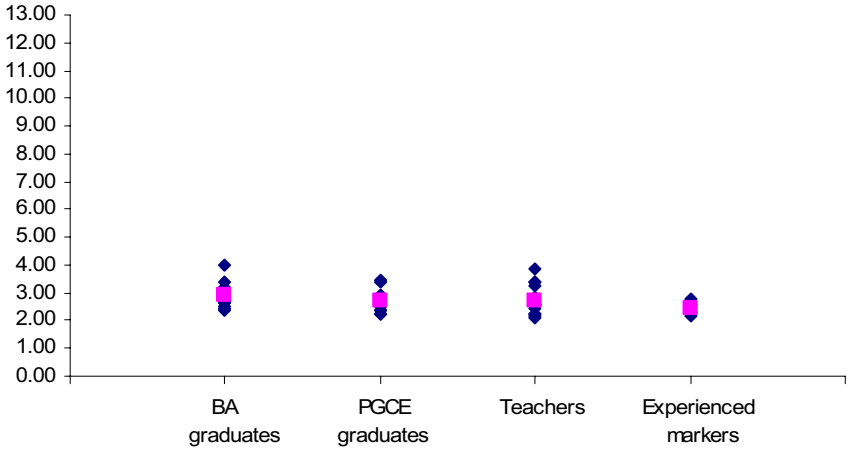
(e) Writing component



(f) Shakespeare reading task



(g) Shakespeare writing task



To test for differences between the size of the groups' absolute mark differences, repeated measures analyses of variances were conducted using the same procedure described in Section 3.1. No significant between-subject effects were found (the results are given in full in Appendix 4). This suggests that the accuracy of the groups, though differing in the amount of variability in the groups, was at a similar level on all aspects of the test. Thus, when compared to the national standard no differences between the groups emerged. There were more and less accurate markers in each group with no group emerging as more or less accurate than any other.

3.6 Defining acceptable marking standards

a. Mark differences at first sample stage

In live marking, a measure of absolute mark difference on ten scripts between the marker and his or her team leader is used as a measure of accuracy against which markers are judged whether to continue marking or be further assessed for marking accuracy. The measure used is the sum of the absolute mark differences captured at strand level for the two writing tasks, at task level for the Shakespeare reading task and at component level for the reading component (see Section 2.2 for definitions of strand, task, component and paper). Figure 6 gives an example of the sample script record form completed by team leaders for each marker in their team to capture the mark differences. The absolute mark difference is calculated by summing the differences for the separate writing strands and reading tasks irrespective of the direction of the difference. This summed absolute mark difference is different to that described as the English test absolute mark difference in Section 3.5 because it aggregates every difference found in each aspect of the test, which results in a more inflated measure than the absolute difference between the two total scores, which will inevitably mask deviations that cancel each other out at strand or component level. In 2003, if a single pupil had a summed absolute mark difference of 10 or more, or if the summed absolute mark difference of ten pupils was more than 66, a marker would have been required to mark more sample scripts to assess his or her acceptability.

Figure 6: Sample script record form with examples of calculations

	Pupil name	*Pupil 1*			*Pupil 2*		
		M	S	D	M	S	D
Longer writing task	SSP	*5*	*5*	-	*1*	*1*	-
	TSO	*4*	*4*	-	*1*	*1*	-
	CE	*6*	*4*	*+2*	*1*	*2*	*-1*
Shorter writing task	SSPTO	*4*	*4*	-	*3*	*3*	-
	CE	*6*	*5*	*+1*	*2*	*3*	*-1*
	Spelling	*3*	*4*	*-1*	*3*	*2*	*+1*
Reading	Reading paper	*17*	*16*	*+1*	*16*	*17*	*-1*
	Shakespeare reading task	*10*	*10*	-	*6*	*6*	-
	AMD			*5*			*4*
	Script total	*55*	*52*	*+3*	*33*	*35*	*-2*

M = marker's mark
S = supervisor's mark
D = difference

The first sample summed absolute mark differences were calculated in the study and are given in Table 13 summarised across the marker types. The first sample scripts in the study yielded absolute mark differences over twice the size deemed a minimum acceptable standard for all marker types. The findings related to 'marks on' and 'marks off' scripts discussed in Section 1 should be brought to bear. The method for standardising markers in the live operation is a 'marks on' method where the team leaders can see the marks given by the markers, but in the study, the method used was a 'marks off' method because the scripts were not re-marked, but instead compared to the Lead Chief Marker's marks. Given the large differences between 'marks on' and 'marks off' mark differences, it should not be surprising that the study yielded substantially higher absolute mark differences than seen in the live marking. Tests of homogeneity of variance for the summed absolute mark differences did not reveal any significant differences between the groups. Similarly, a one-way analysis of variance to test for group differences in the mean summed absolute mark difference did not reveal any significant differences. Full results for both of these tests are given in Appendix 4. This suggests that although the BA graduates appeared more variable and had a higher mean summed absolute mark difference, these differences were not significantly different to any of the other groups.

Table 13: Mean absolute mark differences for the ten sample scripts

	Number	Mean	Standard deviation
BA graduates	170	133.06	28.02
PGCE graduates	150	126.80	16.89
Teachers	140	123.50	17.36
Experienced markers	90	121.67	17.62

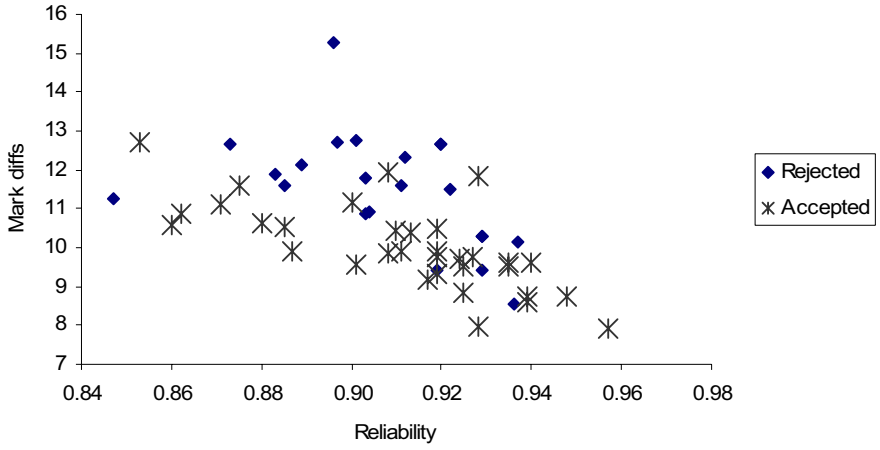
In live marking, markers are rated according to the size of their summed mark difference over the sample. If it is too large, they are required to complete further sample scripts. If they are consistently too far from their team leader, they are asked to not mark any further. To see what would have happened if markers had been stopped from marking at the first sample stage based on large mark differences, a hypothetical limit of 130 was applied to their mark differences. (This limit was arbitrarily selected because it is double that used in live marking which uses a 'marks on' approach to standardisation whereas the study used a 'marks off' approach.) Twenty-two markers would have been excluded, some from every group: eight BA graduates, five PGCE graduates, six teachers and three experienced markers. The absolute mark differences across the entire allocation, as described in Section 3.5 and used as a measure of accuracy, were plotted against the markers' reliability measures in Figure 7a and against percentage same level in Figure 7b. The excluded markers are distinguished from the included markers to highlight who they are.

Both Figures 7a and 7b show that this method of weeding out poor markers is not efficient. Some excluded markers redeemed themselves and achieved mark differences, percentage same level and reliability estimates as good as included markers, suggesting their exclusion was premature. On the other hand, some excluded markers maintained high mark differences and lower reliability and percentage same estimates, suggesting their high first sample mark differences were a good prediction of their future marking. With regard to the

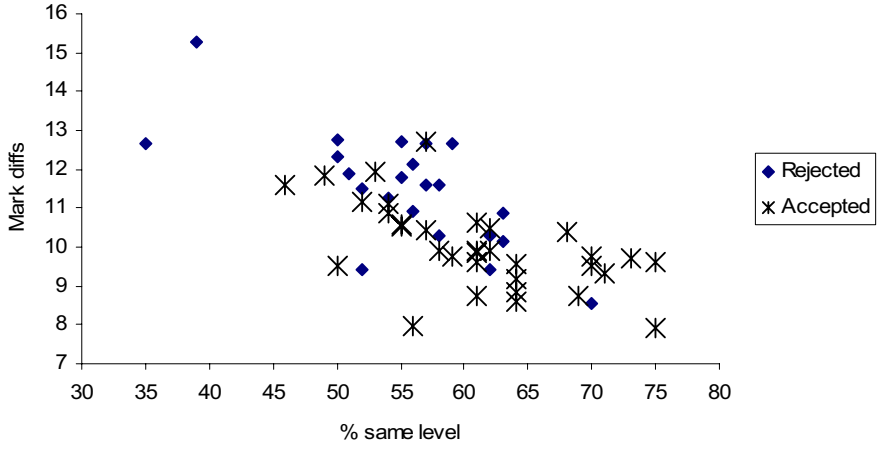
included markers, some were not spotted at the initial sample stage and ended up with low reliability and percentage same estimates, suggesting the initial exclusion process did not detect them. And alternatively, other included markers went on to yield respectable reliability and percentage same estimates, suggesting they were correctly included. That some markers could be wrongly excluded and others wrongly included suggests that the use of a mark difference to exclude markers at the first sample stage on ten scripts may not be the most effective method of weeding out less capable markers.

Figure 7: Hypothetically excluded and included markers:

7a. Reliability estimates against mark differences on entire allocation



7b. Percentage same level against mark differences on entire allocation



b. Limits for reliability and percentage same levels

An alternative approach to determining acceptable marking is to apply hypothetical limits on the reliability and percentage same level estimates. The limit for reliability came from Newton (1996) and a generous cut-off of percentage came from Powers and Kubota (1998) as follows:

Total test	reliability greater than or equal to 0.81 and 50% same or more
Reading paper	reliability greater than or equal to 0.85 and 50% same or more
Writing paper	reliability greater than or equal to 0.74 and 50% same or more

Five markers would be excluded if the limits were applied to the total test scores: two BA graduates, one PGCE graduate, one teacher and one experienced marker. Four markers would be excluded if the limits were applied to the reading paper scores: one of the same BA graduates, the same teacher and two experienced markers, one also excluded by the total test scores. For the writing paper, 29 markers would be excluded: 11 BA graduates (65%), eight PGCE graduates (50%), five teachers (33%) and three experienced markers (33%). So, using these arbitrary limits, markers from all four groups would not be acceptable markers, particularly not for the writing elements. If those rejected from the writing paper were excluded from the entire test, the baby would be thrown out with the bathwater with regards to the reading elements. This suggests that either different limits of acceptability would need to be applied to the writing elements or that certain markers should be targeted for particular papers. However, as a means of determining the quality of markers in live marking, this approach is flawed because it would require much photocopying and duplicated marking effort.

3.7 Clerical errors

The clerical error rates are reported in two sections: script errors and marksheet errors.

a. Script errors

The errors reported here follow closely those recorded in AQA's mainstream examination specifications. Every instance of an error was recorded, not just those that would have resulted in a level change. The rates are not comparable to the rates indicated in the introduction, which are instances of schools requesting a check because the error jeopardised a pupil's final level. The emphasis of these results is on the differences between the marker types. Table 14 shows the error rate by marker group expressed as a mean number per marker. The 17 instances of a marker not marking all of the work turned out to be errors associated with photocopying rather than unmarked work. Pages had been missed out or the photocopied work was too difficult to read, which were errors on the part of the AQA and not the markers.

Table 14: Mean number of script errors per marker reported by marker type

	I. Not all work marked	II. Within script: question total missing	III. Within script: addition error in question total	IV. Front: question total missing	V. Front: wrong question total carried over	VI. Front column total added up wrongly
BA graduates	0.35	0.06	1.65	0.35	0.71	1.29
PGCE graduates	0.25	3.75	2.19	2.56	0.75	3.62
Teachers	0.27	0.93	2.27	1.13	0.40	2.00
Experienced markers	0.33	0.22	1.56	0.44	0.67	2.22
Lead Chief Marker	0.00	0.00	0.00	0.00	0.00	3.00
Raw number of errors	17	77	111	68	36	133

There were only 77 instances where a marker failed to total the questions within the reading component and 60 of these came from two PGCE graduates (column II), which accounts for the relatively high error rate in the PGCE graduate row. This error would normally be spotted at the first phase sample, but because team leaders were not employed, there was not the same attention to detail that there is in live marking.

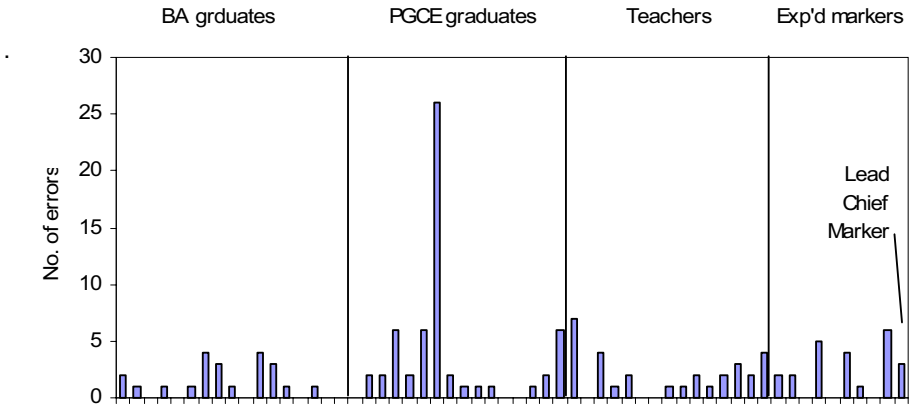
Over half of all markers made at least one addition error inside the reading scripts (column III); and 23 markers did not make any at all. There was little difference between the mean error rates by marker type, but the experienced markers were let down by one marker who notched up 11 errors alone while six of the nine made no errors of this type.

There were only 68 instances of the total mark being left off the front cover over all the components. One of the PGCE markers failed to write it on 29 scripts, which accounts for the high mean error rate for this group in column IV.

The incorrect transfer of marks from the inside of the booklet to the front cover was a rare occurrence, only 36 instances in total (column V). Nearly 60% of the markers did not make this type of error. No marker made more than three errors of this nature in their allocation.

The most common script error was totalling the component marks. Three-quarters of the markers made at least one error in totalling the marks on the front of a script, as shown in Figure 8. Viewed as marker types, the PGCE graduates fared the worst, because one marker in particular made far more errors than any other increasing the mean error rate for the group.

Figure 8: Number of incorrect totalling on script errors made by each marker



Overall, the clerical errors in the scripts were rare. Particular markers made errors, rather than particular marker types, which suggests that errors made on the scripts could be dealt with by team leaders at first and second phase sampling stages.

b. Marksheet errors

Marksheet errors, on the other hand, were not rare. The two most prolific errors made were the totalling of marks and the conversion of marks to levels and they were made by almost all markers. The other types of errors tended to have been made by one or two markers only.

Errors made totalling marks on the marksheets were made by almost all markers: only five markers did not make any at all. Across all markers, 317 errors of this nature were made out of a total number of 17,400 additions calculated on the marksheets, a rate of 1.8%. The mean error rates per marker are given in Table 15, further broken down by marker type. No marker type stands out as being particularly more accurate than any other, as can also be seen in Figure 9, though the Lead Chief Marker fared very well by only making two addition errors.

Table 15: Mean number of totalling and converting errors per marker by marker type

	C. Incorrect totalling of marks on marksheet	F. Incorrect conversion of marks to levels
BA graduates	6.41	22.88
PGCE graduates	4.87	8.67
Teachers	6.21	27.93
Experienced markers	5.11	3.56
Lead Chief Marker	2.00	0.00
Total	317	942

The conversion of marks to levels was calculated incorrectly 942 times across all of the marksheets out of 17,400 conversions, a rate of 5.4%. This was by far the most frequent type of error. Table 15 shows that BA graduates and teachers had the highest error rate which was as a result of two markers in each of the two groups making a large number of errors. A

look at the marksheets indicated that the markers in many cases appeared to have confused the reading and writing levels and applied the level from the wrong skill. However, in some other cases, it is difficult to see from where they got the level they applied. The experienced markers made far fewer conversion errors compared to the other types of markers. This suggests that prior experience of marking contributed to the correct conversion of marks to levels rather than teaching experience. The other markers would presumably improve with experience or additional training. In the questionnaire responses, about 30% of the markers, of all types, felt that clerical responsibilities and marksheet completion could have been covered better in the training.

Figure 9: Number of incorrect totalling of marks on marksheets per marker

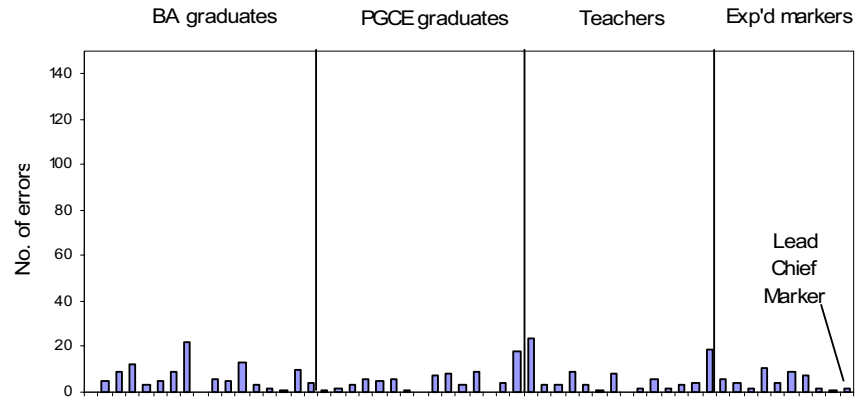
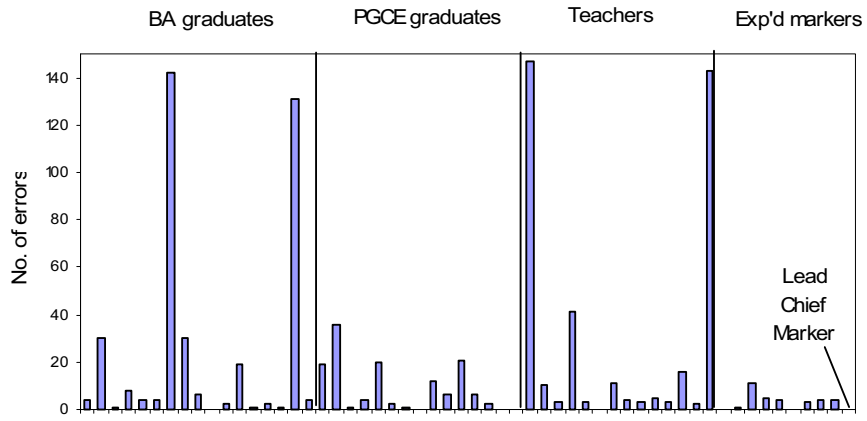


Figure 10: Number of incorrect conversions of marks to levels on marksheets per marker



Other types of errors tended to have been made by one or two markers. The mean error rates by marker types are given in Table 16, but in each case, the mean rate masks errors made only by a handful of markers.

One or more errors, either on scripts or marksheets, were made by every marker. Particular markers tended to make errors rather than marker types. There is no reason to suggest that markers with less teaching experience would make any more errors than those with more teaching experience. It appears to be an individual characteristic.

Table 16: Mean number of other types of marksheet error per marker by marker type

	A. Total given when marks incomplete	B. Total blank when marks given	D. Level given when marks incomplete	E. Level blank when marks given	G. Marks missing	H. Level written in wrong column
BA graduates	0.00	0.06	0.24	17.88	0.06	0.00
PGCE graduates	0.00	0.00	0.00	20.00	0.00	0.00
Teachers	0.07	7.14	0.36	0.29	0.00	7.14
Experienced markers	0.00	0.00	0.00	0.00	0.00	0.00
Lead Chief Marker	0.00	0.00	0.00	0.00	0.00	0.00
Total number	1	101	9	608	1	100

3.8 Feedback from markers

Of the 57 questionnaires sent, 52 were returned completed. The results are reported in sections relevant to the questions asked.

a. Reasons for taking part in study

The markers were asked to give as full an account as possible to explain why they applied to take part in the study. Some markers had more than one reason and in all 102 reasons were given. The most common two reasons were for the money and for the transfer of training to teaching. These reasons were given by markers of all types. As a way of learning about marking and what is expected of Year 9 students was given as a reason by many teachers and graduates, and some experienced markers also noted it was a way to find out about the changes in the test. Financial reasons were given by many markers, but the experience the study afforded was a much more common reason. One marker, a BA graduate, had an article published in the national press describing her experience in the study which suggested journalistic ambitions not alluded to in the questionnaire (*The Guardian*, 2003). Participation in the study, then, was for some a means to experience marking on a small manageable, non-high stakes scale. Setting up small introductory marking sessions for new potential markers could be a way of attracting more markers in the future.

b. Preparation for marking

The markers were sent the training and administration materials before the first training day. They were required to work through several exercises prior to the training days. They were asked whether they thought these materials were adequate for their intended purpose. Almost all of the markers, between 92% and 100%, found the various materials adequate and only six of them rated one or more of them as less than adequate. There was no consensus between these six as to the ones which were less than adequate.

The markers were asked whether they had had enough time to prepare for the training days. Two markers did not receive their materials in time and so clearly they did not. Eight others said they felt that they had not been given enough time to work through the materials, but the majority, 84%, felt the time had been sufficient. Interestingly, three of the eight who felt pressed for time were experienced markers. They all said that they took about twice the amount of time recommended in the Handbook to complete the training exercises. Nine markers commented that some aspect of the materials was confusing or difficult to follow

because of incorrect page numbers. Yet, on the other hand, ten markers said they found the materials straightforward and clear.

c. Training, standardisation and further samples

The two training days covered all aspects of the marking process. The markers were asked to rate how well the various aspects were covered. The mark schemes and commentaries for all components and tasks were rated as being covered well in the training by 95% of the markers. The introduction to the Key Stage 3 English test was rated as covered well by 89% of the respondents and a similar proportion felt the same way about the discussion of the key marking points. The marking of the standardisation scripts could have been covered better according to seven (13%) respondents. Nine markers (17%) said that the further sampling procedures could have been covered better and unsurprisingly five of them were BA graduates who had had no previous marking experience at all. Similarly, 17% of the markers thought the administration and deadlines could have been better covered in the training. Four of these nine markers were teachers. Completion of the marksheets and clerical responsibilities were felt to be poorly covered by about 30% of the respondents. The topic thought to be covered the least well in the training by two-thirds of the markers was borderline checking.

Specific comments about the training were split, half were positive and half were negative. While 15 markers commented that the training was too rushed or not deep enough, another 15 commented that it was thorough and well prepared.

After the training days, the markers were given standardisation scripts to mark which were sent to the Team Supervisors for checking and comment. The majority of markers, 88%, thought this was an adequate way to be trained to mark. Furthermore, 94% of them were happy with the support they received from their Team Supervisors. The 6% who were not happy, three markers, were either teachers or experienced markers who may have had higher expectations or held strong views on their marking accuracy.

Of all of the markers who responded to the questionnaire, 46 said they made changes to their marking based on the feedback from the ten further sample scripts. This is an indication that they took the Lead Chief Marker's marks into consideration and indeed in several cases it is possible to see on the marksheets that the markers altered their marks for scripts prior to and including those on the third marksheet. About 50% of the markers reported that they were in contact with their team supervisors after receiving their further sample feedback. Interestingly, only 25% of the BA graduates contacted their team supervisor while over 60% of the teachers and PGCE graduates were in contact. This again is an indication that as a method of standardisation, it was taken seriously by the markers. Indeed, 66% of them felt it was an adequate way of being trained to mark. Compared to the 88% answering positively with regard to the use of standardisation scripts as a method of being trained to mark, there was some slippage. The comments on the further sample reveal no consensus of opinion. Five markers, for example, felt that using ten scripts was not enough to reveal a pattern of leniency or severity in their marking compared to the Lead Chief Marker. Five other markers felt there was not enough time to go back and re-mark their scripts in light of the feedback, which is an issue even in live marking. Two experienced markers who thought the method

less than adequate felt that more guidance was needed on how to review previously marked work in the light of the further sample commentaries and marks. It is understandable that markers used to the mediation process with their Team Leader would find the method used in the study difficult to adjust to. On the other hand, three markers new to marking, said they found the method very useful.

d. Would you mark again?

The markers were asked if they would mark Key Stage 3 English if the opportunity arose. The majority, 63%, said they would, 11% said they would not and the remaining 25% were undecided. The pattern of responses within each group of markers was not similar as seen in Table 17. Only two of the experienced markers thought they would mark again, while six others were undecided. The BA and PGCE graduates were more enthusiastic to mark again. Of the six who said they would not mark again, three mentioned that marking was too time-consuming or stressful. The reasons given by the 13 undecided markers fell into two camps: two graduates noted they were unsure about the future and two experienced markers noted the marking had become too complex and not as interesting as before. The teachers who were undecided did not give particular reasons.

Table 17: Responses to the question 'Would you mark Key Stage 3 English again?'

	Yes	No	Possibly	Total
BA graduates	14	1	1	16
PGCE graduates	11	1	2	14
Teachers	6	3	4	13
Experienced markers	2	1	6	9
Total	33	6	13	52

e. Benefits

To see how marking could be made attractive to new markers, it is worthwhile looking at the markers' responses to the question about how their involvement in the marking study benefited them. BA graduates noted that the experience gave them a general insight into marking, Key Stage 3 or the education system. Their comments were not specific on how the experience could be used in the future, whereas the PGCE graduates, teachers and some of the experienced markers noted how the experience would directly impact on their current or future work with students. Many of them noted particularly that the knowledge gained about how the examination works would be useful for their teaching and preparing students for the tests.

4. DISCUSSION

The results overall did not suggest an overwhelming difference in marking accuracy and reliability achieved by any of the different marker types in the study. That there were some differences in important, but the main finding was of no difference between the groups with regard to accuracy measured by the absolute mark differences from the Lead Chief Marker.

It could be argued that the experienced markers were not viable as a control group, which is why there were no differences between the groups. There is no prominent reason to

conclude that they were not representative of experienced markers on the evidence of their past marking experience. They had marked before and would have met the criteria to mark in 2003. The markers were all A*, A or B grade markers, except for one who was a C grade marker, which indicates that they were respected and valued markers. 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, they simply attend the training along with everyone else. The training in the study, that all markers underwent, was the same as that used in the live marking, conducted by the same trainers in one of the same locations used earlier in the year. According to their questionnaire responses, and all of the experienced markers returned questionnaires, the main motivations for taking part were to get experience in the new 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. These responses suggest motivation based on a serious professional interest. It may be of relevance that only two of the nine said they would mark again. An indication of future involvement would not necessarily have any effect on how they marked during the study. Whilst motivation may set them apart from the other marker groups, there is no overt reason why they would not be representative of experienced markers.

The groups were equally accurate, or equally inaccurate, depending on one's point of view, indicating that teaching experience was not a contributing factor to marking accuracy. There were differences between the groups' variance measures of the absolute mark differences, but these differences were not indicative of statistical differences in the size of the mark differences. This suggests 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) on 'marks off' scripts, suggesting the mark differences were comparable in size to those found in studies of experienced markers.

No differences between the groups were also found in marking reliability, as defined by agreement between the levels assigned to a pupil by a marker compared to those assigned by the national standard. Again, the implication is that teaching experience did not make any difference. It is interesting that the proportion of disagreements was roughly 40% as estimated by Wiliam (2001), and it mirrors the misclassification rate amongst experienced markers that Powers and Kubota (1998) saw in their study which similarly used a six point scale. The other measure of marking reliability, the correlation coefficient of marks from a marker and the Lead Chief Marker, indicated similar levels of reliability on the reading and writing components. Lower levels amongst the two graduate groups at test level and on the Shakespeare reading task were found, as they were also found amongst the BA graduates on the Shakespeare writing task. These findings suggest that teaching experience was a contributing factor to higher marking reliability in the Shakespeare tasks. Familiarity with this type of exercise may have made the difference or it may have been caused by reluctance, or lack of confidence on the part of these markers to tackle the Shakespeare tasks but not other tasks, or it may be caused by something inherently difficult about the tasks. Reference to the literature (Murphy, 1978 and 1979, 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 the tasks,

components and papers of this examination, as noted by Newton (2003). The comparisons of the reliability estimates in the study were relative, and what was absent was a comparison to an agreed acceptable degree of reliability for the tasks. 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. However, the utility of correlation coefficients to define marker reliability is nonetheless called into question.

Marking reliability expressed as a coefficient of correlation gives an indication on the agreement in rank ordering between the two markers, but this is rendered meaningless as a measure of reliability if two markers are perfectly correlated but differ consistently by ten marks. Clearly, pupils marked by the two markers would receive levels that differed if no adjustment of marks was carried out to take account of the difference in severity. The study showed an absence of a clear linear relationship between markers' correlation measures and their agreement in level rates: higher correlation measures were not necessarily associated with higher agreement rates. The relationship in the study appeared more complex, though it was not investigated in more depth. The findings, perhaps, provide fuel to the dialogue between Newton (2003) and Wiliam (2003) on what constitutes reliable marking and how it should be defined. From the schools' point of view, the proportion of pupils awarded their 'correct' level is a closer approximation of the reliability of marking because pupils' levels determine future action. Whilst marking reviews allow queries on results to be investigated, there is no measurement of marking reliability with due marking adjustments before results are issued.

The amount of variability in the mark differences and the number of misclassified levels suggest there is variability in marking accuracy in live marking. One way accuracy is judged in live marking is through the calculation of absolute mark differences compared to team leaders during the first and second phase sampling. This is the criterion used to judge marking quality. The hypothetical rejection of markers at the first sample explored in Section 3.6 suggested that some markers were rejected who would have achieved marking reliability levels (as defined by percentage same level) similar to accepted markers. Another way accuracy is judged in live marking is by the number of requests for a marking review, but since only the number of pupils who have their levels changed is recorded, other instances of inaccurate marking are not observed. The marking accuracy measured by this study may be one of the first glimpses at the extent to which markers vary in Key Stage 3 English.

Another finding was that of no difference between the relative leniency of the four groups on all aspects of the test except the composite reading paper, where the experienced markers were more lenient than the two groups with teaching experience, but similar to the BA graduates. The same was not found in the paper's two components, the reading component and the Shakespeare reading task. This finding echoes that of Ruth and Murphy (1988) who found a similar difference in lenience between experienced markers and trainee teachers. Teaching experience not yet informed by marking experience, for some reason, contributed to markers assigning fewer marks for these tasks. It should be noted that there was a large amount of variation between the markers within each group, suggesting lenient and severe markers appeared in all groups. The marks from the BA graduates were more variable than those from other markers, but with regard to severity or leniency as a group, they were no

different from the experienced markers: individual BA graduates may differ significantly from the experienced markers with regard to lenience, but as a group, they did not. Operationally, this would suggest that some BA graduates would be capable of an acceptable quality of marking, but their marking would need to be monitored to exclude overly lenient or, more likely, severe markers.

It could be argued that the findings were a result of using the 'marks off' approach to standardisation in that it was less effective in some way and caused the lack of differences between markers and that the use of the 'marks on' approach may have resulted in a different outcome. The literature suggests that it is just as effective as using no standardisation at all (Baird, Bell and Greatorex, 2003) and as effective as the 'marks on' approach (Meadows, *in preparation*). The process in the study still involved professional contact and advice between the team supervisors and the markers, as would the 'marks on' approach. The outcomes of the two approaches differ because the 'marks on' approach yields much lower mark differences than if the scripts are marked blind, indicating that prior knowledge of a marker's judgement affects the team leader's judgement (Murphy, 1978), arguably by informing it and enabling the team leader to support the marker's original decision (Massey and Foulkes, 1994). Yet, it is not simply a matter of size, but one of credibility. It is harder to accept a mark difference of 132 over ten scripts than one of 66 because the former suggests gross inaccuracies in the marks for the individual pupil. However, the former may conceal a tacit amount of inaccuracy because of the mediation effect that prior knowledge of a marker's mark brings. Indeed, senior examiners use both 'marks on' and 'marks off' mark differences to determine the advice they would give to an examiner in accordance with AQA procedure, suggesting they and the examiners are able to differentiate what the mark difference means depending on its origin (Meadows, *in preparation*). While the 'marks on' approach would have seen lower absolute mark differences in the sample, there is little evidence to suggest it would have resulted in different study outcomes.

Another measure of accuracy was taken on the errors made in administration. The results suggest that errors tended to be made by certain individuals rather than types of markers, but overall, experienced markers were more accurate. Teaching experience was not a factor that explained these differences, but rather marking experience. A longer time spent during the training days may result in fewer errors, but clerical errors are likely to keep occurring in the absence of uniform systematic checks.

A 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 Lead Chief Marker's marks is evidence that the markers appreciated the differences between pupils' performance and were not assigning marks with abandon. Also, the relatively small mark differences suggest that markers applied themselves to the task. For these reasons, it is argued that a study effect was unlikely.

5. CONCLUSIONS

The study found no overwhelming differences between markers who had differing amounts of teaching experience. Although the BA graduates indicated tendencies to be more extreme in their marking, they were still capable of comparative levels of quality to experienced teachers and markers. The criterion for markers to have three years of teaching experience appears to be unnecessary and could be relaxed to allow English graduates to mark Key Stage 3 English. The implication is that the training, standardisation and mark scheme are sufficiently rigorous to be effective in preparing non-teaching personnel. Whether this is the case for other subjects at Key Stage 3 or at other Key Stage levels needs to be explored, though it is likely that less subjective subjects would also lend themselves to be adequately marked by non-teachers.

The use of correlation coefficients as a measure of marking reliability is called into question. Higher coefficients are not necessarily an indication of a high proportion of the correct assignation of Key Stage level. Their usefulness needs to be further explored in relation to marking reliability expressed as a measure of the accuracy of the final outcome for the pupil.

The proportion of misclassifications observed in the study is a cause for concern if, as suggested, it mirrors that unobserved in live marking. However, no precedence exists for the accuracy of Key Stage 3 English marking and it is difficult to put this finding into any context. It should be pointed out that practically, there is no absolutely correct level for any pupil because it is an impossibility to have every pupil marked by exactly the same standard. A theoretical 'true score' exists and any attempt at marking a pupil's work will include an element of measurement error obscuring that 'true score'. The task of the External Marking Agency is to minimise the error through the use of standardisation and review procedures. One major difference between Key Stage marking and that of other mainstream marking is the absence of any marker adjustment to account for overly lenient or severe marking. Given the importance attached to the results of Key Stage 3 tests, further investigations into acceptable levels of marking reliability would be welcomed. The recommended removal of the necessity to be an experienced teacher to mark Key Stage 3 English should hopefully provide an increased pool of markers interested in participating in not only live marking but also further studies.

Lucy Royal-Dawson
Senior Research Officer, AQA
30th January 2004

6. REFERENCES

AQA (2003) *Administration File for Supervisors Key Stage 3 English.* Produced in association with Qualifications and Curriculum Authority

Baird, J., Greatorex, J. and Bell, J.F. (2003) *What Makes Marking Reliable? Experiments with UK Examinations.* AQA Research Committee Paper RC/217

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

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

The Guardian (2003) *On your marks* by E. Morcom published on 2nd September 2003 accessed at http://education.guardian.co.uk/egweekly/story/0,5500,1033474,00.html

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

Massey, A. and Foulkes, J. (1993) *Audit of the 1993 KS3 Science National Test Pilot and the Concept of Quasi-reconciliation.* Evaluation and Research in Education v8 n3 p119-132

Meadows, M. (*in preparation*) *The Use of 'Live' Versus Photocopied Scripts in the First Phase Sample of Marking Standardisation.* In draft. For submission to the AQA Research Committee in June 2004

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

Murphy, R.J.L. (1979) *Removing the Marks from Examination Scripts Before Re-marking Them: Does It Make Any Difference?* British Journal of Educational Psychology v49 n1 p73-78

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

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

Pinot de Moira, A. and Davies, C. (2002) *Clerical Errors in Marking: New Specification A Levels, AS, VCE & GNVQ Examinations, Plus SEG Legacy Syllabuses Summer 2002 Examinations.* AQA Research Committee *Paper* RC/192

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

Royal-Dawson, L. (2003) *Study Proposal: Key Stage 3 English Markers Study.* AQA internal document

Ruth, L. and Murphy, S. (1988) *Designing Writing Tasks for the Assessment of Writing.* Norwood, New Jersey: Ablex Publishing Corporation

Shaw, S. (2002) *The effect of training and standardisation on rater judgement and inter-rater reliability.* University of Cambridge Local Examinations Syndicate Research Notes Issue 8

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

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

7. ACKNOWLEDGEMENTS

The implementation of Key Stage 3 English Marker study was made successful by the cooperation and support of many different parties, both from outside and inside AQA. I would like to give my thanks and appreciation for the following support from the following people:

- For assisting with the recruitment, despatches and script control, Anita Cooper and Dawn Kerby.
- For rounding up interest and participants amongst BA graduates and PGCE graduates about to depart for their summer vacations: Moira Brown of College of St Matthew and St Mark, Plymouth; Debra Myhill of Exeter University; Dr Terence Clifford Amos, Jennifer Jackson and Paul Skinner of Canterbury Christchurch University College; Isabel Pain and Steven Matthews of Oxford Brookes University; Linda Fursland and Tim Middleton of Bath Spa University; Gill Payne of University of

Plymouth (Exmouth Campus); Andrew Green and Professor Dominic Head of Brunel University; Marelin Orr-Ewing and Ruth Sharpe of University of the West of England.

- For insights into the marking operation of Key Stage 3 English and support and interest in the progress of the study, Ruth Goddard.
- For overseeing the progress and implementation of the study, Stephen Rigby.
- For managing the standardisation of the markers, Linda Sheppard.
- For the training days and professional insights, John Green and Moira Brown.
- For advice on statistical matters, Mudhaffar Al-Bayatti, Jo-Anne Baird, Debra Dhillon and Lesley Meyer.

APPENDIX 1

2003 Key Stage 3 English Level Thresholds

Writing

Level	Mark range
4	6 – 13
5	14 – 22
6	23 – 32
7	33 – 50

Reading

Level	Mark range
4	10 – 15
5	16 – 26
6	27 – 33
7	34 – 50

English Overall

Level	Mark range
N	0 – 10
3	11 – 15
4	16 – 29
5	30 – 49
6	50 – 66
7	67 – 100

APPENDIX 2A

Script Checking Error Form for KS3 English Marker Study

Marker's Number: _____ **Bundle Number:** _____ of 5 **Date:** _____

Script Checker's Name: _____

Paper	No. in packet	No. checked	Type of error Write candidate number against error type	Total no. of errors	Photocopying errors. Give candidate no.
Reading			Not all the work marked Within script: Q total missing Within script: addition error in Q total Front: Q total missing Front: wrong Q total carried over Front: column total added wrongly Front: borderline check ticked		
Writing			Not all the work marked Front: Q total missing Front: column total added wrongly Within script: inappropriate mark or notes		
Shakespeare			Not all the work marked Front: Q total missing Front: column total added wrongly Within script: inappropriate mark or notes		
			TOTALS		

PLEASE WRITE OVERLEAF ANY NOTES MADE BY MARKER ON FRONT-COVER

Script checker's signature: _____

APPENDIX 2B

Marksheet Error Form

Marker Number …… **Checker ……**

Type of error	Tally	Total
Total written when marks incomplete		
Total blank when all marks present		
Total incorrect		
Level written when marks incomplete		
Level blank when all marks present		
Level incorrect		
Other error 1 …….. (state type of error)		
Other error 2 …… (state type of error)		
Total		

APPENDIX 3

ASSESSMENT and QUALIFICATIONS ALLIANCE

KEY STAGE 3 ENGLISH MARKER STUDY

Questionnaire for markers

Your name: ……..…….…….. Marker id number: ……

Before the training meetings

1. Was the *Administration File for markers and supervisors* adequate for its purpose?

	Inadequate			**Adequate**
	1	**2**	**3**	**4**
Please tick one box ✓				

2. Were the training modules sent to you in advance of the training meetings adequate for their purpose?

	Inadequate			**Adequate**
Please tick one box per row ✓	**1**	**2**	**3**	**4**
Module 1 – The Reading Paper				
Module 2 – The Shakespeare Reading Tasks				
Module 3 – Dealing with Administration				
Module 4 – The Longer Writing Task				
Module 5 – The Shorter Writing Task				

3. Did you have sufficient time after the training materials were delivered to use them as intended?

	Insufficient time			**Sufficient time**
	1	**2**	**3**	**4**
Please tick one box ✓				

4. Do you have any comments on the *Administration File*, the training modules or the time allowed for the training exercises?

At the training meetings

5. How well do you consider each of the following areas was covered at the training meetings?

	Not very well			**Very well**
Please tick one box per row ✓	**1**	**2**	**3**	**4**
Introduction to the KS 3 English test				
Discussion of key marking points				
Completion of marksheets				
Mark schemes & commentaries for the reading paper				
Mark schemes & commentaries for the writing paper				
Mark schemes & commentaries for the Shakespeare reading task				
Mark scheme & commentaries for the Shakespeare writing task				
Marking the standardisation scripts				
Further sample procedures				
Clerical responsibilities				
Administration and deadlines				
Borderline checking				

6. Do you have any comments on the training you received?

After the training meetings

7. Was the use of the standardisation scripts an adequate method to train you to mark?

	Inadequate			**Adequate**
	1	**2**	**3**	**4**
Please tick one box ✓				

8. Was the support you received from your Team Supervisor in relation to the standardisation scripts sufficient?

	Insufficient			**Sufficient**
	1	**2**	**3**	**4**
Please tick one box ✓				

9. Did you have any contact with your Team Supervisor after receiving your mark differences and commentaries on the 10 further sample scripts?

	Yes	**No**
Please tick one box ✓		

10. Do you have any comments on the support you received from your Team Supervisor?

---

11. Did you make any changes to your marking based on the feedback from the 10 further sample scripts?

	Yes	**No**
Please tick one box ✓		

12. Was the use of the further sample mark differences and commentaries an adequate method to train you to mark?

	Inadequate			**Adequate**
	1	**2**	**3**	**4**
Please tick one box ✓				

13. Do you have any comments on the further sample mark differences and commentaries?

---

14. Did you develop a marking routine? For example, mark to set hours, set a target number to reach, etc.

If yes, please describe the routine you developed.

If no, please describe how you marked.

---

15. What benefits, if any, has marking KS3 English scripts in this study brought to you?

Before and After Marking

16. Why did you take part in the study? There may be several reasons, so please give as full an account as possible.

---

17. Would you mark again in the near future for KS3 if the opportunity arose? (Your response is not binding nor does it guarantee future work.)

For those who had <u>not</u> done any external marking before:

18. What was your impression of marking external examinations before you tried it?

---

19. Has the experience of marking altered your impression of marking? If so, please explain how.

Any other comments

20. Do you have any other comments about the marking procedures, administration, potential benefits or anything else in relation to the Key Stage 3 English marker study?

When you have completed this questionnaire, please return it to Lucy Royal-Dawson, AQA, Stag Hill House, Guildford GU2 7XJ or use the pre-paid envelope provided.

Thank you very much.

Please note that your responses in this questionnaire will be treated in accordance with AQA's Research Code of Practice (available upon request), which ensures the identity of respondents is kept confidential.

APPENDIX 4 – Results from Section 3

Section 3.1 Raw marks – tests of homogeneity of variance

The test statistic used is:

$$F = s_L^2 / s_S^2 \sim F_{0.025}(N_L\text{-}1, N_S\text{-}1)$$

where s_L^2 and s_S^2 represent the larger and smaller of the two variances respectively.

	BA graduates vs PGCE graduates	BA graduates vs teachers	BA graduates vs experienced markers	PGCE graduates vs teachers	PGCE graduates vs experienced markers	Teachers vs experienced markers
English test	1.069 *p*=0.09	**1.117** *p*=0.01	1.105 *p*=0.04	1.045 *p*=0.20	1.034 *p*=0.29	1.011 *p*=0.42
Reading paper	1.075 *p*=0.07	1.087 *p*=0.05	**1.131** *p*=0.02	1.012 *p*=0.41	1.052 *p*=0.20	1.040 *p*=0.26
Writing paper	1.046 *p*=0.18	**1.110** *p*=0.02	1.039 *p*=0.25	1.061 *p*=0.12	1.007 *p*=0.45	1.068 *p*=0.14
Reading component	1.095 *p*=0.03	1.024 *p*=0.32	1.113 *p*=0.03	1.069 *p*=0.10	1.016 *p*=0.39	1.086 *p*=0.08
Writing component	1.054 *p*=0.14	1.088 *p*=0.05	1.008 *p*=0.44	1.032 *p*=0.27	1.046 *p*=0.22	1.079 *p*=0.10
Shakespeare reading task	1.011 *p*=0.41	1.077 *p*=0.07	1.060 *p*=0.16	1.089 *p*=0.05	1.072 *p*=0.12	1.015 *p*=0.40
Shakespeare writing task	1.029 *p*=0.28	**1.106** *p*=0.02	1.056 *p*=0.17	1.074 *p*=0.08	1.026 *p*=0.33	1.047 *p*=0.22

where $F > F_{0.025}(df_L, df_S)$ F is shown **in bold**

Section 3.1 - Raw marks - repeated measures analyses of variance results

PUPIL are the within-subject variables; MTYPE is the between-subject factor.

Source	Adjustment for non-sphericity	Type III Sum of Squares	df	Mean Square	F	Sig.
English test						
Tests of Within-Subjects Effects						
PUPIL	Greenhouse-Geisser	1015144.54	14.70	69049.28	368.84	0.00
	Huynh-Feldt	1015144.54	27.52	36891.20	368.84	0.00
PUPIL * MTYPE	Greenhouse-Geisser	10498.83	44.11	238.04	1.27	0.12
	Huynh-Feldt	10498.83	82.55	127.18	1.27	0.06
Error(PUPIL)	Greenhouse-Geisser	99081.00	529.26	187.21		
	Huynh-Feldt	99081.00	990.62	100.02		
Tests of Between-Subjects Effects						
Intercept		6466782.00	1	6466782.001	5595.11	0.00
MTYPE		9044.92	3	3014.972341	2.61	0.07
Error		41608.51	36	1155.791846		
Reading paper						
Tests of Within-Subjects Effects						
PUPIL	Greenhouse-Geisser	297791.74	21.23	14028.54	414.46	0.00
	Huynh-Feldt	297791.74	53.04	5614.80	414.46	0.00
PUPIL * MTYPE	Greenhouse-Geisser	2676.21	63.68	42.02	1.24	0.10
	Huynh-Feldt	2676.21	159.11	16.82	1.24	0.03
Error(PUPIL)	Greenhouse-Geisser	27303.00	806.65	33.85		
	Huynh-Feldt	27303.00	2015.40	13.55		
Tests of Between-Subjects Effects						
Intercept		1847103.61	1	1847103.61	7202.02	0.00
MTYPE		2272.64	3	757.55	2.95	**0.04**
Error		9745.87	38	256.47		
Writing paper						
Tests of Within-Subjects Effects						
PUPIL	Greenhouse-Geisser	292207.94	16.84	17350.18	176.78	0.00
	Huynh-Feldt	292207.94	29.69	9841.01	176.78	0.00
PUPIL * MTYPE	Greenhouse-Geisser	5620.20	50.53	111.24	1.13	0.25
	Huynh-Feldt	5620.20	89.08	63.09	1.13	0.19
Error(PUPIL)	Greenhouse-Geisser	72728.57	741.04	98.14		
	Huynh-Feldt	72728.57	1306.49	55.67		
Tests of Between-Subjects Effects						
Intercept		1692511.50	1	1692511.50	2804.58	0.00
MTYPE		2730.18	3	910.06	1.51	0.23
Error		26553.20	44	603.48		

Source	Adjustment for non-sphericity	Type III Sum of Squares	df	Mean Square	F	Sig.
Reading component						
Tests of Within-Subjects Effects						
PUPIL	Greenhouse-Geisser	136518.34	22.47	6076.54	377.20	0.00
	Huynh-Feldt	136518.34	54.44	2507.62	377.20	0.00
PUPIL * MTYPE	Greenhouse-Geisser	1321.07	67.40	19.60	1.22	0.12
	Huynh-Feldt	1321.07	163.32	8.09	1.22	0.04
Error(PUPIL)	Greenhouse-Geisser	14838.76	921.13	16.11		
	Huynh-Feldt	14838.76	2232.10	6.65		
Tests of Between-Subjects Effects						
Intercept		912167.56	1	912167.56	7922.62	0.00
MTYPE		635.61	3	211.87	1.84	0.15
Error		4720.52	41	115.13		
Writing component						
Tests of Within-Subjects Effects						
PUPIL	Greenhouse-Geisser	97282.63	19.45	5002.02	96.63	0.00
	Huynh-Feldt	97282.63	35.93	2707.66	96.63	0.00
PUPIL * MTYPE	Greenhouse-Geisser	3579.00	58.35	61.34	1.19	0.17
	Huynh-Feldt	3579.00	107.79	33.20	1.19	0.10
Error(PUPIL)	Greenhouse-Geisser	47316.53	914.09	51.76		
	Huynh-Feldt	47316.53	1688.65	28.02		
Tests of Between-Subjects Effects						
Intercept		550031.66	1	550031.66	1989.83	0.00
MTYPE		957.04	3	319.01	1.15	0.34
Error		12991.82	47	276.42		
Shakespeare reading component						
Tests of Within-Subjects Effects						
PUPIL	Greenhouse-Geisser	49152.85	23.48	2093.08	162.52	0.00
	Huynh-Feldt	49152.85	52.06	944.18	162.52	0.00
PUPIL * MTYPE	Greenhouse-Geisser	1363.61	70.45	19.36	1.50	0.01
	Huynh-Feldt	1363.61	156.18	8.73	1.50	0.00
Error(PUPIL)	Greenhouse-Geisser	13912.00	1080.24	12.88		
	Huynh-Feldt	13912.00	2394.70	5.81		
Tests of Between-Subjects Effects						
Intercept		216824.00	1	216824.00	2898.31	0.00
MTYPE		419.18	3	139.73	1.87	0.15
Error		3441.29	46	74.81		
Shakespeare writing component						
Tests of Within-Subjects Effects						
PUPIL	Greenhouse-Geisser	70521.56	22.15	3184.00	164.13	0.00
	Huynh-Feldt	70521.56	45.37	1554.25	164.13	0.00
PUPIL * MTYPE	Greenhouse-Geisser	1373.16	66.45	20.67	1.07	0.34
	Huynh-Feldt	1373.16	136.12	10.09	1.07	0.29
Error(PUPIL)	Greenhouse-Geisser	20193.85	1040.99	19.40		
	Huynh-Feldt	20193.85	2132.55	9.47		
Tests of Between-Subjects Effects						
Intercept		351581.60	1	351581.60	2627.38	0.00
MTYPE		526.69	3	175.56	1.31	0.28
Error		6289.28	47	133.81		

Section 3.2 Reliability estimates – tests of homogeneity of variance

(df1, df2)	BA graduates vs PGCE graduates (17, 16)	BA graduates vs teachers (17,15)	BA graduates vs experienced markers (17,9)	PGCE graduates vs teachers (16,15)	PGCE graduates vs experienced markers (16,9)	Teachers vs experienced markers (15,9)
English test	2.065 p=0.08	1.847 p=0.12	3.379 p=0.03	1.118 p=0.41	1.636 p=0.23	1.829 p=0.18
Reading paper	**4.259** p<0.01	1.694 p=0.16	**5.258** ***p*=0.01**	2.515 p=0.04	1.235 p=0.39	3.105 p=0.04
Writing paper	**2.993** ***p*=0.02**	2.438 p=0.04	2.752 p=0.06	1.228 p=0.34	1.087 p=0.42	1.129 p=0.44
Reading component	1.874 p=0.11	1.410 p=0.25	2.211 p=0.11	1.330 p=0.29	1.180 p=0.41	1.569 p=0.25
Writing component	**3.085** ***p*=0.01**	1.224 p=0.35	1.607 p=0.24	2.520 p=0.04	1.920 p=0.12	1.313 p=0.35
Shakespeare reading task	1.247 p=0.33	2.682 p=0.03	**3.956** ***p*=0.02**	2.151 p=0.07	3.173 p=0.04	1.475 p=0.28
Shakespeare writing task	1.881 p=0.11	1.393 p=0.26	2.494 p=0.08	1.351 p=0.28	1.326 p=0.34	1.791 p=0.19

where F > $F_{0.025}(df_L, df_S)$ F is shown **in bold**

Section 3.2 Reliability estimates – one-way analysis of variance

One-way analysis of variance conducted on the transformed correlation coefficients for the marks for each marker and the Lead Chief Marker. The Welch procedure was used to account for unequal sample sizes in the test and lack of homogeneity of variance in some cases (Howell, 1992).

	One-way analysis of variance						ANOVA using the Welch procedure			
	Source	Sum of Squares	df	Mean Square	F	Sig.	Welch statistic	df1	df2	Sig.
English test	Between Groups	0.2267	3	0.0756	4.66	0.01	4.64	3	27.67	**0.009**
	Within Groups	0.8598	53	0.0162						
	Total	1.0865	56							
Reading paper	Between Groups	0.0993	3	0.0331	2.29	0.09	2.90	3	27.34	0.053
	Within Groups	0.7670	53	0.0145						
	Total	0.8663	56							
Writing paper	Between Groups	0.1475	3	0.0492	3.24	0.03	2.93	3	26.42	0.052
	Within Groups	0.8033	53	0.0152						
	Total	0.9508	56							
Reading component	Between Groups	0.0140	3	0.0047	0.39	0.76	0.44	3	26.91	0.729
	Within Groups	0.6273	53	0.0118						
	Total	0.6413	56							
Writing component	Between Groups	0.0829	3	0.0276	1.51	0.22	1.22	3	25.21	0.321
	Within Groups	0.9702	53	0.0183						
	Total	1.0531	56							
Shakespeare reading task	Between Groups	0.3762	3	0.1254	5.39	0.003	5.96	3	28.04	**0.003**
	Within Groups	1.2325	53	0.0233						
	Total	1.6087	56							
Shakespeare writing task	Between Groups	0.1631	3	0.0544	3.29	0.03	3.08	3	27.21	**0.044**
	Within Groups	0.8752	53	0.0165						
	Total	1.0383	56							

Section 3.3 Percentage same level – one-way analysis of variance

	One-way analysis of variance						ANOVA using the Welch procedure			
	Source	Sum of Squares	df	Mean Square	F	Sig.	Welch statistic	df1	df2	Sig.
English test	Between Groups	169.47	3	56.49	0.81	0.49	0.77	3	25.20	0.52
	Within Groups	3690.04	53	69.62						
	Total	3859.51	56							
Reading paper	Between Groups	250.52	3	83.51	0.77	0.52	0.75	3	24.86	0.53
	Within Groups	5753.62	53	108.56						
	Total	6004.14	56							
Writing paper	Between Groups	101.68	3	33.89	0.60	0.62	0.74	3	26.31	0.54
	Within Groups	2988.04	53	56.38						
	Total	3089.72	56							

Section 3.3 Contingency table for test levels

	3 levels lower	2 levels lower	1 level lower	Same	1 level higher	2 levels higher	3 levels higher	Total
BA graduates	0	23	281	966	361	14	1	1646
PGCE graduates	0	23	246	934	303	14	0	1520
Teachers	0	34	254	916	249	9	0	1462
Exp'd markers	0	6	105	512	240	7	0	870
Total	0	86	886	3328	1153	44	1	5498

Section 3.3 Contingency table for reading paper levels

Counts	3 levels lower	2 levels lower	1 level lower	Same	1 level higher	2 levels higher	3 levels higher	Total
BA graduates	0	2	149	1055	394	32	0	1632
PGCE graduates	0	1	138	1006	344	19	0	1508
Teachers	0	2	117	974	336	20	0	1449
Exp'd markers	0	0	52	525	265	21	0	863
Total	0	5	456	3560	1339	92	0	5452
Percentages								
BA graduates	0.00	0.12	9.13	64.64	24.14	1.96	0.00	
PGCE graduates	0.00	0.07	9.15	66.71	22.81	1.26	0.00	
Teachers	0.00	0.14	8.07	67.22	23.19	1.38	0.00	
Exp'd markers	0.00	0.00	6.03	60.83	30.71	2.43	0.00	

Section 3.3 Contingency table for writing paper levels

Counts	3 levels lower	2 levels lower	1 level lower	Same	1 level higher	2 levels higher	3 levels higher	Total
BA graduates	4	59	377	802	372	47	0	1661
PGCE graduates	5	44	348	798	325	45	0	1565
Teachers	3	48	410	739	241	25	1	1467
Exp'd markers	0	14	162	458	212	30	0	876
Total	12	165	1297	2797	1150	147	1	5569
Percentages								
BA graduates	0.24	3.55	22.70	48.28	22.40	2.83	0.00	
PGCE graduates	0.32	2.81	22.24	50.99	20.77	2.88	0.00	
Teachers	0.20	3.27	27.95	50.37	16.43	1.70	0.07	
Exp'd markers	0.00	1.60	18.49	52.28	24.20	3.42	0.00	

Section 3.5 Absolute mark differences – tests of homogeneity of variance

	BA graduates vs PGCE graduates	BA graduates vs teachers	BA graduates vs experienced markers	PGCE graduates vs teachers	PGCE graduates vs experienced markers	Teachers vs experienced markers
English test	**1.112** ***p*=0.02**	**1.215** ***p*<0.01**	**1.161** ***p*=0.01**	1.093 *p*=0.04	1.045 *p*=0.24	1.047 *p*=0.22
Reading paper	**1.185** ***p*<0.01**	**1.236** ***p*<0.01**	1.049 *p*=0.21	1.043 *p*=0.21	**1.243** ***p*<0.01**	**1.296** ***p*<0.01**
Writing paper	1.021 *p*=0.34	1.028 *p*=0.29	**1.251** ***p*<0.01**	1.007 *p*=0.45	**1.225** ***p*<0.01**	**1.217** ***p*<0.01**
Reading component	1.097 *p*=0.03	1.038 *p*=0.23	**1.285** ***p*<0.01**	**1.138** ***p*=0.01**	**1.171** ***p*<0.01**	**1.333** ***p*<0.01**
Writing component	1.004 *p*=0.47	1.079 *p*=0.06	1.069 *p*=0.13	1.074 *p*=0.08	1.064 *p*=0.15	1.009 *p*=0.44
Shakespeare reading task	**1.331** ***p*<0.01**	**1.377** ***p*<0.01**	**1.239** ***p*<0.01**	1.035 *p*=0.25	1.074 *p*=0.11	1.112 *p*=0.04
Shakespeare writing task	**1.190** ***p*<0.01**	1.002 *p*=0.48	**1.385** ***p*<0.01**	**1.192** ***p*<0.01**	**1.163** ***p*=0.01**	**1.387** ***p*<0.01**

where F > $F_{0.025}(df_L, df_S)$ F is shown **in bold**

Section 3.5 Absolute mark differences – repeated measures analyses of variance

PUPIL are the within-subject variables; MTYPE is the between-subject factor.

English test

Tests of Within-Subjects Effects

Source	Adjustment for non-sphericity	Type III Sum of Squares	df	Mean Square	F	Sig.
PUPIL	Greenhouse-Geisser	18242.87	15.08	1209.59	9.19	0.00
	Huynh-Feldt	18242.87	28.74	634.66	9.19	0.00
PUPIL * MTYPE	Greenhouse-Geisser	8232.62	45.25	181.95	1.38	0.05
	Huynh-Feldt	8232.62	86.23	95.47	1.38	0.01
Error(PUPIL)	Greenhouse-Geisser	71454.12	542.95	131.60		
	Huynh-Feldt	71454.12	1034.79	69.05		

Tests of Between-Subjects Effects

Intercept		160639.49	1	160639.49	929.71	0.00
MTYPE		855.79	3	285.26	1.65	0.19
Error		6220.25	36	172.78		

Reading paper

Tests of Within-Subjects Effects

PUPIL	Greenhouse-Geisser	4653.40	19.50	238.66	9.74	0.00
	Huynh-Feldt	4653.40	44.15	105.39	9.74	0.00
PUPIL * MTYPE	Greenhouse-Geisser	1675.96	58.49	28.65	1.17	0.19
	Huynh-Feldt	1675.96	132.46	12.65	1.17	0.10
Error(PUPIL)	Greenhouse-Geisser	18151.29	740.92	24.50		
	Huynh-Feldt	18151.29	1677.78	10.82		

Tests of Between-Subjects Effects

Intercept		40413.98	1	40413.98	504.20	0.00
MTYPE		642.48	3	214.16	2.67	0.06
Error		3045.89	38	80.15		

Writing paper

Tests of Within-Subjects Effects

PUPIL	Greenhouse-Geisser	20563.99	14.64	1404.17	17.17	0.00
	Huynh-Feldt	20563.99	23.88	861.19	17.17	0.00
PUPIL * MTYPE	Greenhouse-Geisser	4258.28	43.93	96.92	1.19	0.20
	Huynh-Feldt	4258.28	71.64	59.44	1.19	0.15
Error(PUPIL)	Greenhouse-Geisser	52697.32	644.38	81.78		
	Huynh-Feldt	52697.32	1050.66	50.16		

Tests of Between-Subjects Effects

Intercept		118026.31	1	118026.31	1823.64	0.00
MTYPE		277.86	3	92.62	1.43	0.25
Error		2847.69	44	64.72		

Reading component

Tests of Within-Subjects Effects

PUPIL	Greenhouse-Geisser	4186.53	20.63	202.96	15.74	0.00
	Huynh-Feldt	4186.53	45.46	92.09	15.74	0.00
PUPIL * MTYPE	Greenhouse-Geisser	895.49	61.88	14.47	1.12	0.25
	Huynh-Feldt	895.49	136.39	6.57	1.12	0.17
Error(PUPIL)	Greenhouse-Geisser	10905.40	845.70	12.90		
	Huynh-Feldt	10905.40	1863.95	5.85		

Tests of Between-Subjects Effects

Source	Adjustment for non-sphericity	Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept		24595.63	1	24595.63	733.81	0.00
MTYPE		225.69	3	75.23	2.24	0.10
Error		1374.23	41	33.52		

Writing component

Tests of Within-Subjects Effects

PUPIL	Greenhouse-Geisser	8936.30	17.61	507.43	13.44	0.00
	Huynh-Feldt	8936.30	30.49	293.06	13.44	0.00
PUPIL * MTYPE	Greenhouse-Geisser	2372.73	52.83	44.91	1.19	0.17
	Huynh-Feldt	2372.73	91.48	25.94	1.19	0.11
Error(PUPIL)	Greenhouse-Geisser	31243.45	827.71	37.75		
	Huynh-Feldt	31243.45	1433.15	21.80		

Tests of Between-Subjects Effects

Intercept		61596.85	1	61596.85	1233.30	0.00
MTYPE		89.20	3	29.73	0.60	0.62
Error		2347.39	47	49.94		

Shakespeare reading task

Tests of Within-Subjects Effects

PUPIL	Greenhouse-Geisser	1623.98	22.52	72.12	8.75	0.00
	Huynh-Feldt	1623.98	47.86	33.93	8.75	0.00
PUPIL * MTYPE	Greenhouse-Geisser	581.04	67.55	8.60	1.04	0.38
	Huynh-Feldt	581.04	143.57	4.05	1.04	0.35
Error(PUPIL)	Greenhouse-Geisser	8535.58	1035.75	8.24		
	Huynh-Feldt	8535.58	2201.36	3.88		

Tests of Between-Subjects Effects

Intercept		14862.13	1	14862.13	1309.06	0.00
MTYPE		49.93	3	16.64	1.47	0.24
Error		522.25	46	11.35		

Shakespeare writing task

Tests of Within-Subjects Effects

PUPIL	Greenhouse-Geisser	5240.25	16.74	313.11	16.05	0.00
	Huynh-Feldt	5240.25	28.14	186.24	16.05	0.00
PUPIL * MTYPE	Greenhouse-Geisser	1059.57	50.21	21.10	1.08	0.33
	Huynh-Feldt	1059.57	84.41	12.55	1.08	0.29
Error(PUPIL)	Greenhouse-Geisser	15344.30	786.60	19.51		
	Huynh-Feldt	15344.30	1322.45	11.60		

Tests of Between-Subjects Effects

Intercept		34510.34	1	34510.34	2180.65	0.00
MTYPE		105.26	3	35.09	2.22	0.10
Error		743.81	47	15.83		

Section 3.6 - summed absolute mark differences from the first sample

One-way analysis of variance

	One-way analysis of variance						ANOVA using the Welch procedure			
	Source	Sum of Squares	df	Mean Square	F	Sig.	Welch statistic	df1	df2	Sig.
English test	Between Groups	1027.82	3	342.618	.750	.527	0.572	3	25.23	0.638
	Within Groups	22374.18	49	456.62						
	Total	23402.00	52							

Tests of homogeneity of variance

	BA graduates vs PGCE graduates	BA graduates vs teachers	BA graduates vs Experienced markers	PGCE graduates vs teachers	PGCE graduates vs experienced markers	Teachers vs experienced markers
English test	2.75 *p*=0.03	2.61 *p*=0.04	2.53 *p*=0.08	1.06 *p*=0.46	1.09 *p*=0.43	1.03 *p*=0.47

where F > $F_{0.025}(df_L, df_S)$ F is shown **in bold**