Examiner Background and the Effect on Marking Reliability

Anne Pinot de Moira
Assessment & Qualifications Alliance

May 2003

This research has been funded by QCA following recommendations made in the independent panel report on maintaining GCE A Level standards published in January 2002

CONTENTS

SUMMARY **3**

INTRODUCTION **5**

BACKGROUND 5
AWARDING BODY PRACTICES 5
MARK-REMARK RELIABILITY 6
BIAS AND CONTEXT EFFECTS 6
EXAMINER CHARACTERISTICS 7

METHODOLOGY **9**

THE DATA 9
THE MODELS 10
VARIABLE FORMULATION 11

FINDINGS **12**

MODEL 1 – THE PERCENTAGE DIFFERENCE MODEL 12
MODEL 2 – THE ABSOLUTE PERCENTAGE DIFFERENCE MODEL 14
MODEL 3 – THE ADJUSTMENT MODEL 17
MODEL 4 – THE EXAMINER PERFORMANCE RATING MODEL 18
QUESTIONNAIRE TO PRINCIPAL EXAMINERS 20

CONCLUSIONS **24**

RECOMMENDATIONS **25**

REFERENCES **26**

APPENDIX A Variable Formulation 29
APPENDIX B Residual Plots & Model Diagnostics 33
APPENDIX C Questionnaire to Principal Examiners 37

SUMMARY

Arising from recommendations of the independent panel report on maintaining GCE A Level standards (Baker, McGraw, & Lord Sutherland of Houndwood, January 2002), this report discusses a study of the background of examiners and the marks they give. Even though there is little published literature which relates reliability to examiner characteristics, the presented work is set in the context of existing marking reliability research.

Data from a sample of 21 AQA A2 units, marked by 356 examiners in summer 2002 has been analysed by fitting four multilevel models. Each model considers a different aspect of marking reliability as represented by four statistical measures: the difference between senior examiner and assistant examiner mark; the absolute difference between senior examiner and assistant examiner mark; the probability of a numerical adjustment having been made to the assistant examiner's marks and the examiner performance rating. Unit, examiner, centre and candidate level independent variables are included where they explain a significant amount of variation in the dependent variable.

The study identifies no link between personal characteristics and marking reliability. Evidence suggests that reliability is more closely related to features of an examiner's allocation and the idiosyncrasies of individual subjects. The models produce some evidence to support the argument that the work of more able candidates is harder to mark, as is the work of candidates from independent and selective establishments. Questionnaire responses from Principal Examiners shed some light on possible reasons for the observed centre type differences. Recommendations are made for future research in the area with a view to gaining a greater understanding of the influences on marking reliability and to using this understanding to operational advantage.

Keywords: *Marking reliability; accuracy; examiner background; multilevel model.*

INTRODUCTION

Background

In 2001, QCA commissioned a panel of independent experts to assess the quality assurance measures in place to maintain GCE A Level standards. In the context of the investigation, standards were defined by the demands of the specifications, the assessment practices in place and the performance of candidates. The remit of the panel was to

> '..... review the overall quality assurance arrangements for GCE A level against best international practice.'
>
> (Baker, McGraw, & Lord Sutherland of Houndwood, January 2002)

Although the panel's findings were generally favourable, one of the key recommendations was that QCA should assume 'a more proactive research stance' with a view to informing future debate on GCE A Level standards. Among the areas considered essential for research was quality of marking and, following publication of the report, QCA issued tenders for work in the field. This report arises from the tender for research into the backgrounds of examiners and the marks they give. It attempts to identify factors which might allow awarding bodies to predict those assistant examiners who are likely to be most efficient and those who are likely to require additional training or monitoring.

Despite considerable literature covering the issues surrounding marking reliability, there is little research into examiner backgrounds. Findings from the study are therefore presented in the context of existing marking reliability publications.

Awarding Body Practices

Measures to assess the quality of marking for general and vocational qualifications offered in England, Wales and Northern Ireland are firmly embedded in a code of practice produced in consultation between the Qualification and Curriculum Authority (QCA), the Curriculum and Assessment Authority for Wales (ACCAC) and the Council for Curriculum, Examinations and Assessment (CCEA) (QCA, ACCAC, & CCEA, 2002). All awarding bodies have agreed to implement the code of practice in full and, with this commitment, have accepted recommendations for good practice in the standardisation of marking for external assessment. Pertinent to matters of marking reliability are the following excerpts from the code of practice (QCA et al., 2002), labelled with the appropriate paragraph numbers.

> ' 51. All examiners should have relevant experience in the subject area ……
> 54. All examiners must satisfactorily complete all aspects of the standardisation process ……
> 57. The awarding body must ensure that all examiners have a well-founded and common understanding of the requirements of the mark scheme and can apply them reliably ……
> 60. Immediately after the standardisation meeting of examiners, assistant examiners must mark fully a sample of scripts ….. and forward them to a more senior examiner ……
> 62. Examiners must not proceed to finalise any marking until they have received clearance from the relevant senior examiners ……
> 63. The continuing marking of all examiners must be monitored by the appropriate senior examiner ……'

However, control over quality of marking comes not only from the regulatory authorities but also from

the research community and awarding bodies themselves. They have invested considerable energy into considering matters of marking reliability and consistency. The scope of projects undertaken in the past has been widespread and, while some findings have been disseminated through the usual network of journal publications, a great deal have been used in a purely operational capacity to inform and improve awarding body practices. Marking reliability literature which is not currently in the public domain is discussed alongside published work in order to set the current study into the context of both existing research and awarding body practices.

Mark-Remark Reliability

Over the years examiner reliability has been continually monitored using mark-remark experiments to evaluate the effect of the differing conditions under which marking is performed. In an assessment of the reliability of marking in eight GCE examinations, Murphy (1978) observed that there were several factors affecting examiner accuracy. He listed the number of components within an examination, the subject matter and the type of question as contributing to the levels of marking reliability. Although it is a simple mathematical inevitability that accuracy should increase as the number of examined elements increases, the effect that subject matter and question type have upon reliability is less clear. The empirical evidence presented by Murphy (1978) suggested closely defined questions are more reliably marked than free-response questions. Indeed, a meta-analysis of 29 mark-remark experiments, conducted by the Associated Examining Board between 1976 and 1980 showed that, while marking reliability was high in all subject areas under consideration, it was highest for mathematics and lowest for English. Furthermore, a study conducted by Hartog & Rhodes (1935) concluded that an analytical approach to marking provided greater reliability than an impressionist approach.

While many of the early mark-remark analyses drew conclusions from relatively naïve comparisons of correlation coefficients, later experiments imposed conditions on examiners and analysed the impact of these conditions using more robust statistical tools. For example, in a study of GCE Business Studies, Baird & Pinot de Moira (1997) made changes to the mark scheme in order to evaluate its influence on the marking process. Baird, Greatorex, & Bell (2002) performed further research considering the effect of increasing the detail in the mark scheme and introducing different styles of standardisation meeting. Neither analysis supported the hypothesis that marking reliability was affected by the different conditions applied.

Bias and Context Effects

Experiments have also considered factors over which test administrators have less control. Baird (1998), for example, discussed the implications of gender bias and handwriting style on marks awarded. While she found no evidence of marking bias in the subjects under consideration, other commentators have shown that examiners do appear to introduce bias into student assessment. Archer & McCarthy (1988) reviewed literature in the field of bias and concluded that marking reliability might be affected by the sex, social class and physical attractiveness of a candidate. Furthermore, they suggested the halo – or contrast – effect of knowledge gained from prior assessment activity might influence opinion of current work. This effect was pursued by Spear (1996) who asked teachers to judge a number of features of an essay, having previously shown them other work by the same candidate. Findings suggested that the teachers were prejudiced by the ancillary information they had from the candidate. In Spear (1997), the experiment was extended to consider the effect that order of presentation of work has on the mark awarded. This too was shown to bias marking.

The halo effect has clear implications for the marking of externally assessed general and vocational qualifications. Within AQA, examiners are instructed to mark one centre at a time and, as far as possible, mark in numerical sequence of centre and candidate numbers. While these instructions

make attempts to remove any element of choice from the marking sequence, neither centre number nor candidate number are allocated randomly. Centre number is assigned regionally and candidate number is assigned by the centre. There is some evidence to suggest that contrast or halo effects are at their greatest at the beginning of the marking exercise and that reading a good range of scripts in advance might minimise the problems experienced (Hughes, Keeling, & Tuck, 1980). The standardisation meeting required by the code of practice (QCA et al., 2002) may facilitate such familiarisation. Equally the community of practice built up amongst examiners may enhance joint understanding (Baird et al., 2002). Furthermore Shaw (2002) speculated 'the mark scheme, comprising a set of detailed and explicit descriptors, engenders a standardising effect even in the absence of a formalised training programme'. Indeed, in considering marking reliability in a longitudinal context, Pinot de Moira, Massey, Baird, & Morrissy (2002) found there was only minor change in the relative leniency or severity of examiners over the period of marking summer 2000 GCE English scripts. This work supported conclusions drawn by Lunz & O'Neill (1997) who showed that although individual judges vary in their level of leniency, the leniency of most judges remains internally consistent over time, in spite of retraining.

Examiner Characteristics

The personal characteristics of individual examiners, such as those discussed by Lunz & O'Neill (1997), have been the subject of less research. Nevertheless, quality assurance measures in place for examiner recruitment tend to assume that good practice revolves around experienced examiners. Several high profile bodies within the United Kingdom clearly regard experience as essential to marking accuracy.

> 'To qualify for consideration [as an examiner], you must normally be teaching the subject concerned (or a related subject) with 3 years' experience of preparing candidates for National Qualifications courses, or 3 years' assessment experience in tertiary education.'
>
> (Scottish Qualifications Authority, n.d.)

> 'Institutions should consider developing and employing criteria to support the appointment of external examiners, which will normally make reference to: appropriate levels of academic and/or professional expertise and experience in relation to the relevant subject area and assessment'
>
> (The Quality Assurance Agency for Higher Education, January 2000)

The code of practice (QCA et al., 2002) also demands that examiners must have relevant experience in the subject and, although there is no explicit discussion of the nature of this experience, some awarding bodies suggest that three years' teaching in a relevant subject area is desirable.

With the proliferation of examining and the introduction of computer-based assessment, the search for a definition of 'relevant experience' has taken on a new importance. Examiners are in short supply and new technology will eventually provide the facility for individual items within an examination to be marked separately, possibly by clerical staff. Indeed, at the National Foundation for Educational Research (NFER), an online marking pilot for Year 7 Progress Tests in mathematics and English considered, among other issues, the effect of using unskilled and semi-skilled examiners to mark specifically chosen items (Whetton & Newton, 2002). The findings suggested that, with some intervention by supervisors, this strategy could be technically effective. A similar, though less extensive, pilot study was undertaken by AQA in the marking of GCE Chemistry (Fowles, 2002). Although the focus of the study was the reliability of e-marking in comparison with conventional marking, the results suggested that, with sensibly chosen items, clerical marking could provide a

reliable alternative to the use of skilled resources.

In order to inform the development of their Online Scoring Network, the Educational Testing Service have also invested research energies into assessing reliability of marking by an appropriately trained inexperienced group of individuals (Powers & Kubota, 1998a). In a comparison of this group with experienced examiners, they concluded that 'there were few significant relations between background and accuracy', although all their inexperienced examiners were recruited with the minimum of a graduate qualification. Powers & Kubota (1998b) extended this research and collected logical reasoning scores[1] for those involved in the study. The results suggested a possible link between logical reasoning and marking accuracy.

Attempts to link personality traits with marking performance have also been made by Branthwaite, Trueman, & Berrisford (1981) and Pal (1986). However, the small scale nature of these studies, and somewhat ambiguous personality indices, precludes sensible interpretation of the effect that examiner characteristics can exert on marking reliability. Further published work in this field seems to focus on the influence of examiner personality when the examiner is also the test administrator.

The majority of marking carried out for general qualifications within England, Wales and Northern Ireland is performed by examiners marking scripts from candidates with whom they have had no personal contact. The continual monitoring of these examiners is designed to identify doubt over reliability of marking. This monitoring is necessary because the recruitment practice of employing examiners with relevant experience is not sufficient to guarantee marking accuracy. Although the monitoring task is supported by statistical evidence, it is largely impressionistic and is therefore time consuming and imprecise. During 1998 and 1999, the awarding sub-group of the Joint Council for General Qualifications Technical Issues Sub-Committee (JCGQ TISC) discussed the matter of formalising techniques for identification of examiners for whom there remained lingering doubt at the end of the marking period. This discussion gave rise to analyses of examiner backgrounds, centre statistics and candidate characteristics with a view to isolating outlying examiner performances (Meyer, 2000; Bell, 2000). The models were fitted to live data for operational purposes only and, rather than trying to establish *a priori* the features which might contribute to accurate marking, they simply attempted to explain all justifiable differences between candidate – and by implication examiner – performances. Although reasonably successful with the identification of errant examiners, the models provided no evidence to suggest examiner level variables influenced marking accuracy, nor did they explain all variation in the data.

In order to predict, at the recruitment stage, those examiners who will produce accurate marking and those who will require additional training or support, the analyses presented within this report attempt to identify from the available data the qualities of a reliable examiner. They also consider possible differences between the subjects being examined. The research has been designed as a pilot to test a methodology which might be used in the future and therefore the conclusions are presented with this remit in mind.

[1] As determined from 25 logical reasoning questions selected from the Educational Testing Service's Graduate Record Examinations General Test.

METHODOLOGY

The Data

All data used in the evaluation of marking reliability were taken from AQA A2 units examined in summer 2002 (Table 1). The units included covered a range of subject areas and data were available from most examiners involved in the marking. Within the sample, there were 5,942 candidates from 1,510 centres, marked by 356 examiners. The work of each candidate included in the dataset had been remarked by a senior examiner as part of the second phase sample marking exercise[2]. Therefore the unadjusted assistant examiner mark and the mark awarded by a senior examiner were available for use in the analysis.

TABLE 1 AQA A2 units included in the analysis

Specification	Unit	Description	Examiners	Centres	Candidates
Computing	CPT4	Processing and Programming Techniques	18	67	315
	CPT5	Advanced Systems Development	16	63	292
Geography A	GGA4	Challenge and Change in the Natural Environment	18	76	291
	GGA5	Challenge and Change in the Human Environment	17	61	303
Geography B	GGB4	Global Change	11	72	205
	GGB5	Synoptic	11	61	218
Information & Communication Technology	ICT4	Information Systems within Organisations	38	97	520
	ICT5	Information: Policy, Strategy and Systems	39	111	516
English Literature A	LTA4	Texts in Time	40	95	648
	LTA6	Reading for Meaning	45	109	717
Mathematics A	MAP1	Pure Mathematics 1	22	64	180
	MAP2	Pure Mathematics 2	14	101	163
	MAP3	Pure Mathematics 3	15	90	184
	MAP6	Pure Mathematics 6	2	80	175
Mathematics B	MBP2	Pure Mathematics 2	9	67	186
	MBP5	Pure Mathematics 5	6	73	177
	MBP6	Pure Mathematics 6	1	55	173
	MBP7	Pure Mathematics 7	1	48	151
Physics A	PA04	Waves, Fields and Nuclear Energy	7	32	103
	PA10	Synoptic	17	55	264
Psychology B	PHB5	Perspectives, Debates and Methods	9	33	161
Total			**356**	**1,510**	**5,942**

[2] The second phase sample marking exercise provides an opportunity for the senior examiner to review an assistant examiner's marking performance. In this phase, 50 scripts are sent to the senior examiner who selects a random sample of 15 to remark. If all the work is deemed satisfactory, no further scripts are considered. If there is any doubt over the marking then a further 10 scripts are remarked. The data collected as part of this exercise are used for verbal feedback to the assistant examiner and to inform future adjustment strategies. Normally this is the last phase in which remarking is completed, although support from the senior examiner continues throughout the marking period. Sometimes, however, there are further post-award checks on an assistant examiner's marking. These further checks may be triggered under a number of circumstances. For example, there may be lingering doubt about the marking of an assistant examiner or, alternatively, there may be evidence to suggest that, for all centres within the examiner's allocation, there is a systematic difference between the centre estimates and the final grades. A small minority of the data points included in the sample for analysis in this study may emanate from these post-award checks. However, no data are included from the first phase sample marking because this phase is part of the examiner standardisation process and, for each of the 10 scripts remarked, the assistant examiner is required to adopt the mark awarded by the senior examiner.

The dataset also included details of candidate sex, candidate age, the grade awarded for the unit, the mean GCSE result attained by the candidate concerned and the number of GCSE results contributing to that mean. For each centre, there were details of the centre type and the number of candidates entered. For each examiner, information available included sex, qualifications, present employment, years since appointment to AQA, examiner rank, whether a marking adjustment had been applied, size of allocation, the grade awarded for marking performance and the mean mark for his or her allocation. Regrettably, the performance rating was not available for years prior to 2002 because the study happened to coincide with the introduction of the new A Level qualification. Although many of the examiners had previous association with AQA, missing prior performance data would have rendered the dataset too small for valid analysis.

The Models

Four separate multilevel models were fitted to the data to assess the effect of examiner background on the reliability of marking.

Model 1 – The Percentage Difference Model

A four level, linear multilevel model with candidate nested within centre, examiner and unit was created. The dependent variable was the percentage difference between assistant examiner mark and senior examiner mark – a negative value denoting severity in the assistant examiner's marking and a positive value denoting leniency.

Model 2 – The Absolute Percentage Difference Model

A four level, linear multilevel model with candidate nested within centre, examiner and unit was created. The dependent variable was the absolute percentage difference between assistant examiner mark and senior examiner mark – a larger value denoting greater discrepancy between the assistant examiner and the senior examiner.

Model 3 – The Adjustment Model

A two level, logistic multilevel model with examiner nested within unit was created. The dependent variable was a binary contrast distinguishing whether an examiner's marking had incurred a numerical adjustment or not. To fit this model, it was necessary to aggregate the dataset so that information was presented at an examiner, not candidate, level. This reduced the size of the dataset from 5,942 observations to 356 observations. Because the 5,942 candidates included in *Model 1* and *Model 2* were only a sample of those marked by each of the examiners, examiner level independent variables describing the centre and candidate composition of each allocation, were created from the full dataset not the sample.[3] The new variables, calculated from the examiners' full allocation, were: percentage of candidates from selective centres; percentage from further education centres; percentage of female candidates; percentage of candidates under 18 years old; percentage of 18 years old and percentage older than 18.

[3] In *Model 1* and *Model 2* the dependent variable is directly related to the candidate and centre level independent variables. In *Model 3* and *Model 4*, the centre and candidate level data are constant for each examiner. For these models, as the centre and candidate level data explain no variation in the dependent variable, a parsimonious approach includes aggregating the data to examiner level. However, it is still important to retain information about examiners' allocations, as it is possible this has some influence over examiner performance. In aggregating the data to examiner level, measures of the characteristics of the candidature can be created. If these measures are derived from the sample of 5,942 candidates, they could be biased by the non-random nature of the sample. For example, if the sample includes 10% of the work marked by an examiner but, by chance, only includes female candidates, an aggregate variable created from the sample would imply that the examiner only marked the work of female candidates. An aggregate variable created from the complete allocation of each examiner might provide a very different, and more accurate, picture. In both these cases the value of the dependent variable for the examiner would remain the same. The implications of sample selection in the creation of robust parameter estimates are discussed further in Pinot de Moira (2002).

Model 4 – The Examiner Performance Rating Model

A two level, linear multilevel model with examiner nested within unit was created. The dependent variable was the examiner performance rating as determined by the senior examiner at the end of the marking period. It was treated as a continuous variable with A=5, B=4, C=3, D=2 and E=1. For the same reasons as those described for *Model 3*, the candidate level data were aggregated to examiner level.

Each model was fitted using MLwiN (Rasbash et al., 2000). *Model 1*, *Model 2* and *Model 4* were determined using an iterative generalised least squares (IGLS) convergence routine in order that the reported deviance statistics were appropriate for making nested model comparisons to gauge improvement in model fit. The non-linear model, *Model 3*, was derived using second order penalised quasilikelihood (PQL) to minimise bias in the parameter estimates.

As all independent variables were deemed to have potential educational significance, each was introduced to the model using a stepwise approach. For *Model 1*, *Model 2* and *Model 4*, it was possible to assess the statistical significance of a variable by considering the difference in deviance upon introducing that variable. For *Model 3*, the statistical significance of the parameter estimate, the predictive efficiency of the model and the variance explained by the model were all considered. In all cases, the final model included only independent variables shown to have a statistically significant fixed effect on the dependent variable. Interaction effects were not investigated exhaustively but were introduced where judged pertinent.

Random parameter estimates associated with the significant fixed effects were then included where appropriate. In some circumstances, by explaining variability between groups within a level, these random effects influenced the parameter estimates and standard errors associated with the fixed effects. The fixed effects were still retained even if the introduction of the random effects rendered them not statistically significant.

Variable Formulation

The independent variables available to control for the effect of examiner background on marking reliability are described in Appendix A. Details of the coding applied and the variable names used within the model are also explained.

FINDINGS

Model 1 – The Percentage Difference Model

There were three independent variables which appeared to have an effect on the difference between mark awarded by an assistant examiner and that awarded by the senior examiner (Table 2). None was an examiner level variable and the model provided no evidence to support a hypothesis that examiner background might impact upon marking reliability. In fact the statistically significant fixed effects described features of the composition of the examiner's allocation.

TABLE 2 Parameter estimates from the multilevel model, modelling the percentage difference between the assistant examiner mark and the senior examiner mark

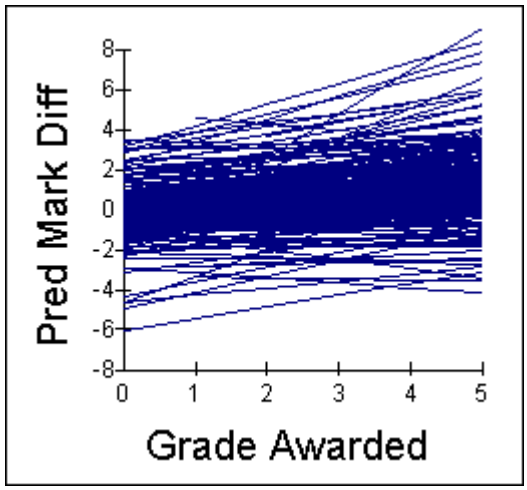
			β	se	p
Fixed Effects	Constant		-0.107	0.185	0.563
	Candidate Grade (Continuous)		**0.244**	**0.039**	**0.000**
	Selective		**-0.748**	**0.248**	**0.003**
	Candidate Mean GCSE Result		**-0.304**	**0.077**	**0.000**
	Candidate Grade * Selective		**0.157**	**0.068**	**0.021**
Random Effects	**Unit Level**	$\sigma^2_{Constant, Constant}$	**0.339**	**0.162**	**0.036**
	Examiner Level	$\sigma^2_{Constant, Constant}$	**2.368**	**0.362**	**0.000**
		$\sigma^2_{Grade, Constant}$	**-0.254**	**0.086**	**0.003**
		$\sigma^2_{Grade, Grade}$	**0.206**	**0.036**	**0.000**
		$\sigma^2_{MeanGCSE, Constant}$	-0.194	0.165	0.240
		$\sigma^2_{MeanGCSE, Grade}$	**-0.246**	**0.059**	**0.000**
		$\sigma^2_{MeanGCSE, MeanGCSE}$	**0.668**	**0.142**	**0.000**
	Centre Level	$\sigma^2_{Constant, Constant}$	**0.596**	**0.141**	**0.000**
	Candidate Level	$\sigma^2_{Constant, Constant}$	**7.041**	**0.168**	**0.000**

Statistically significant effects (α=0.05) emboldened

The first effect to be included in the model was grade awarded to a candidate for the unit under consideration. In order to contextualise this effect, it is necessary to consider the design of the mark scheme which should allow marks to be awarded across the full mark range. Because the mark range is finite, the extent of difference between an assistant examiner and senior examiner mark is more limited for candidates at the extremes of the mark range than in the middle. If it were assumed that the spread of marks awarded by a senior examiner was greater than that by an assistant examiner, the assistant examiner would appear severe at the top end of the mark distribution and lenient at the bottom. The significant independent variable indicating candidate grade suggested the opposite. In other words, the higher a candidate's awarded grade for a unit, the greater the tendency for the marking to be lenient. The raw data confirmed a greater spread of marks by assistant examiners than senior examiners (15.3% compared with 15.1%).

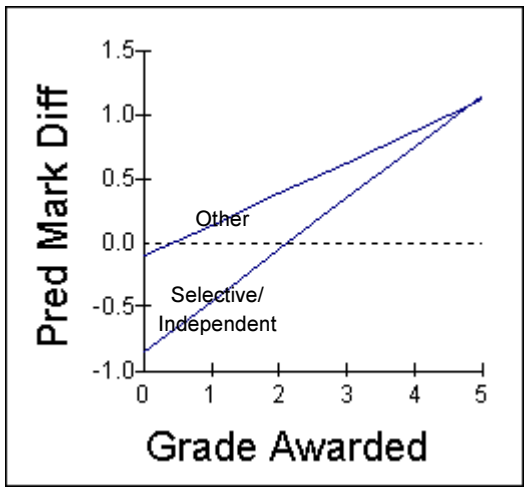
The effect of the grade awarded to a candidate also varied significantly by examiner. Not only was there significant variation in the examiner level intercept ($\sigma^2_{Constant, Constant}$) which described overall leniency or severity, there was also significant slope variation ($\sigma^2_{Grade, Grade}$) and slope/intercept covariance ($\sigma^2_{Grade, Constant}$; Figure 1). Most assistant examiners awarded marks within ± 2% of the senior examiner no matter what the quality of the work they were considering. Notwithstanding this high level of accuracy, the general trend was such that examiners who were most severe to lower ability candidates had the steepest slopes. It was these examiners who appeared most affected by the quality of work they were viewing. Examiners who were relatively lenient in their marking of the lower ability candidates were most likely to produce a consistent marking performance across the whole range of abilities.

FIGURE 1 The examiner level relationship between grade awarded and predicted mark difference



The second fixed effect to be included in the model was a binary contrast denoting whether the candidate was entered through a selective/independent school (1) or through any other educational establishment (0). The interaction between this contrast and the grade awarded to candidates was also significant. Any suggestion that centre type of candidate entry has an effect on the accuracy of marking is disconcerting, particularly in view of the fact that examiners are supposed to know no more about their allocation than the centre number. Compared with all other educational establishments, the marking applied to selective and independent centres appeared to be consistently more severe, although there was evidence that the severity was less pronounced for higher ability candidates (Figure 2). This relative severity did not imply, however, that the marking was consistently less accurate. For candidates awarded a grade D or higher, the marking applied to work from selective or independent centres appeared more accurate. On the other hand, for the lower ability candidates, the marking applied to work from other educational establishments appeared more accurate.

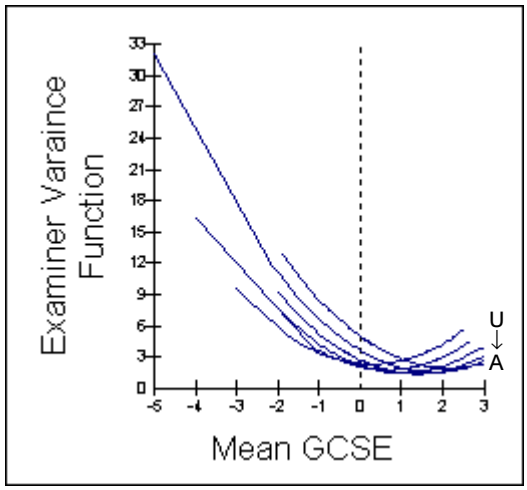
FIGURE 2 The relationship between grade awarded and predicted mark difference dependent upon centre type of entry



Speculation as to why this pattern arose, led to the dispatch of a questionnaire to Principal Examiners involved in the units under consideration (Appendix C). Responses are considered at the end of the section in the light of findings from each of the four models.

The final fixed effect which appeared to have a significant impact on the difference between assistant examiner and senior examiner mark was candidate mean GCSE result. Strangely, despite providing a secondary pseudo measure of ability, this variable had a very different relationship with the dependent variable than that seen for grade awarded. The higher the mean GCSE result, the more severe the marking. Random effects included in the model suggested that there was an examiner level difference in intercept ($\sigma^2_{Constant,\ Constant}$) and slope variance ($\sigma^2_{Grade,\ Grade}$) but there was no intercept/slope covariance ($\sigma^2_{Grade,\ Constant}$). So, whether an examiner had an underlying propensity for leniency or severity, there was no consistent way in which this propensity was affected by the latent ability of the candidate under consideration. There was, however, greater variation in the performance of examiners at the lower end of the candidate ability range (Figure 3). Perhaps this was because A Level examinations are not really designed for candidates with very low mean GCSE results. Therefore, for all but the best examiners, the marking of work completed by candidates with low prior achievement proved more difficult.

FIGURE 3 **Variation in the relationship between assistant examiner and senior examiner mark dependent on the mean GCSE result of the candidate under consideration (graphed separately for each possible grade awarded to candidate)**



Although there was no significant slope variation or intercept/slope covariance at a unit level, there was variation in intercept between units. The difference between assistant examiner and senior examiner mark, therefore, varied by unit. For example, when compared with the senior examiner mark, the marking applied to ICT5, GGB4 and GGA5 tended to be generous, whereas that applied to LTA6, CPT4 and LTA4 tended to be severe. For all other units, the assistant examiner marks lay within $\pm$ 0.25% of the senior examiner marks.

As for all four of the models discussed, residual plots and model diagnostics are presented in Appendix B.

Model 2 – The Absolute Percentage Difference Model

When creating a parsimonious model describing the absolute difference between mark awarded by the assistant examiner and that awarded by the senior examiner, five fixed effects explained significant variation in the dependent variable (Table 3). Four of these five effects described features of the examiner's allocation. The fifth effect was years since appointment to AQA. For every year of experience there was a decrease in the absolute mark difference between assistant and senior

examiner, implying that experience led to an increase in the accuracy of marking. To a certain extent, however, the years of employment are inextricably confounded with marking reliability. Examiners are only retained if their marking continues to be to a high standard. As time proceeds one would expect examiner attrition to concentrate the pool of examiners to those with the most expertise.

TABLE 3 Parameter estimates from the multilevel model, modelling the absolute percentage difference between the assistant examiner mark and the senior examiner mark

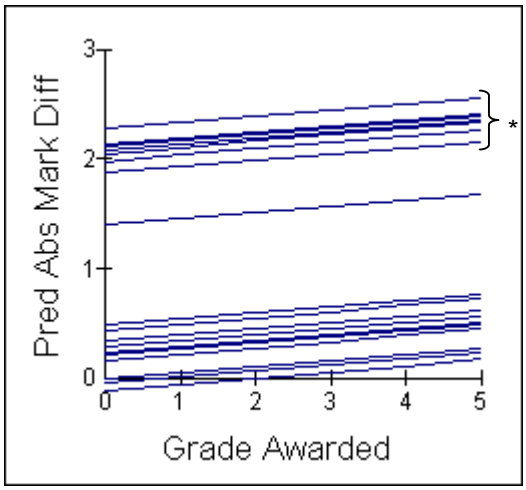
			β	se	p
Fixed Effects	**Constant**		**0.910**	**0.233**	**0.000**
	Candidate Grade (Continuous)		**0.056**	**0.021**	**0.007**
	Years since Appointment		**-0.043**	**0.020**	**0.036**
	Selective		**0.241**	**0.110**	**0.028**
	Allocation		-0.001	0.001	0.245
	Candidate Age		0.021	0.012	0.080
Random Effects	**Unit Level**	$\sigma^2_{Constant, Constant}$	**0.894**	**0.318**	**0.005**
	Examiner Level	$\sigma^2_{Constant, Constant}$	**1.047**	**0.216**	**0.000**
		$\sigma^2_{Grade, Constant}$	**-0.136**	**0.043**	**0.002**
		$\sigma^2_{Grade, Grade}$	**0.029**	**0.012**	**0.014**
		$\sigma^2_{Allocation, Constant}$	**-0.004**	**0.001**	**0.000**
		$\sigma^2_{Allocation, Grade}$	**0.001**	**0.000**	**0.020**
		$\sigma^2_{Allocation, Allocation}$	**0.000**	**0.000**	**0.000**
	Centre Level	$\sigma^2_{Constant, Constant}$	**0.760**	**0.176**	**0.000**
		$\sigma^2_{Grade, Constant}$	**-0.123**	**0.049**	**0.012**
		$\sigma^2_{Grade, Grade}$	**0.053**	**0.017**	**0.002**
	Candidate Level	$\sigma^2_{Constant, Constant}$	**4.338**	**0.103**	**0.000**
		$\sigma^2_{Allocation, Constant}$	**-0.005**	**0.000**	**0.000**
		$\sigma^2_{Allocation, Allocation}$	**0.000**	**0.000**	**0.000**

Statistically significant effects (α=0.05) emboldened

Of the features of examiner's allocation, it was the final grade awarded to the candidate which had the greatest impact on the relationship between assistant examiner and senior examiner mark. The higher the final grade awarded, the less accurate the marking. Whether lenient or severe, examiners appeared to experience more problems with awarding accurate marks to candidates at the top end of the grade scale than to those at the bottom. Though statistically significant, from an operational point of view, the difference in reliability of marking for the higher and lower ability candidates was probably within tolerances (Figure 4).[4] For any given unit, the predicted difference in the dependent variable was less than 0.3% for a candidate awarded grade U compared with a candidate awarded grade A. Indeed there were far greater differences between the units themselves. A clutch of the units under consideration exhibited a very high level of marking reliability, with differences between assistant examiner and senior examiner marks being less than 1%. For others, including LTA4, ICT5, ICT4, GGA4, GGA5, GGB4 and LTA6 the differences were slightly higher but still within acceptable levels (see * on Figure 4).

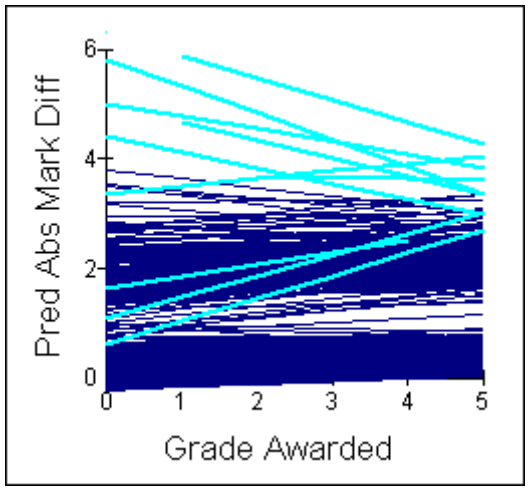
[4] The operational tolerance which triggers further action is normally a difference of greater than 5% of the maximum mark between the mark awarded by the assistant examiner and that awarded by the senior examiner. However, this tolerance is applied to an individual script and, at an examiner level therefore, the aggregate assistant examiner marks and senior examiner marks would be expected to appear better aligned.

FIGURE 4 **The relationship between grade awarded and predicted absolute mark difference dependent upon unit**



Variation in the relationship between the dependent variable and the grade awarded to candidates at examiner and centre level extended beyond differing examiner and centre intercepts. The model suggested that, in general, the more accurate an examiner when marking the lower ability candidates, the more positive the slope describing the relationship between grade awarded and accuracy. However, for many of the examiners included in the study, there was very little difference in accuracy across the full range of candidate abilities. The pattern of converging accuracy towards the top end of the candidate ability range, described by the negative slope/intercept covariance, probably emanated from few examiners. Influence statistics, calculated as a combination of residual and leverage values, measured the impact each examiner had on the random coefficients (Rasbash et al., 2000). The eleven examiners with the largest influence tended to follow a pattern of convergence which was not so evident for the others. These eleven examiners are highlighted in grey on Figure 5.

FIGURE 5 **The examiner level relationship between grade awarded and predicted absolute mark difference (examiners exerting the greatest influence highlighted in grey)**



There were also certain centres which exerted greater influence over the random coefficients. These centres were, on the whole, marked by examiners who themselves had large influence statistics. However, the influential centres did not represent the examiners' complete allocation, suggesting that examiners experienced more problems marking the work submitted by some centres than that submitted by other centres.

As with *Model 1*, the centre type through which a candidate was entered appeared to affect the relationship between assistant and senior examiner mark. The marking applied to scripts from selective or independent centres was slightly less accurate than that applied to scripts from other educational establishments. This effect is discussed alongside the similar *Model 1* findings at the end of the section.

There were two further fixed effects which were significant before any random slope and slope/intercept coefficients were introduced to the model. The first of these was the size of the examiner's script allocation and the second, the age of the candidate. In general, the larger the allocation, the more accurate the marking. Because allocations are made on the basis of competence, the effect of allocation size is, however, of limited operational significance. New examiners have their allocation size limited during the first year of marking. Experienced examiners take on further work throughout the examining period to compensate for examiner shortfall, dropout or remarks.

Interestingly, the evidence presented by *Model 2*, suggested a weak relationship between candidate age and examiner accuracy. The younger the candidate, the closer the mark awarded by the assistant examiner and the senior examiner. It is possible that extricating subject competence from writing maturity may make the task of marking more complex because examiners are normally used to marking work from candidates at the A Level target age.

Model 3 – The Adjustment Model

Within AQA, the decision to adjust the marking of an examiner is taken by the Subject Officer in consultation with a senior member of staff from the Processing Division. To determine the appropriate course of action for each assistant examiner, reference is made to second phase sample marking records, examiner statistics based upon scripts marked to date, historic information and anecdotal evidence from the senior examiner. The decision making process includes an element of subjectivity and it is, perhaps, this subjectivity which renders weak the model designed to predict the probability of an examiner adjustment (Table 4).

TABLE 4 Parameter estimates from the multilevel model, modelling the probability that an examiner is adjusted

		β	se	p
Fixed Effects	**Constant**	**-2.157**	**0.321**	**0.000**
	Proportion from Selective Centres	**2.825**	**0.955**	**0.003**
Random Effect	Unit Level $\sigma^2_{Constant,\ Constant}$	1.079	0.577	0.061

Statistically significant effects (α=0.05) emboldened

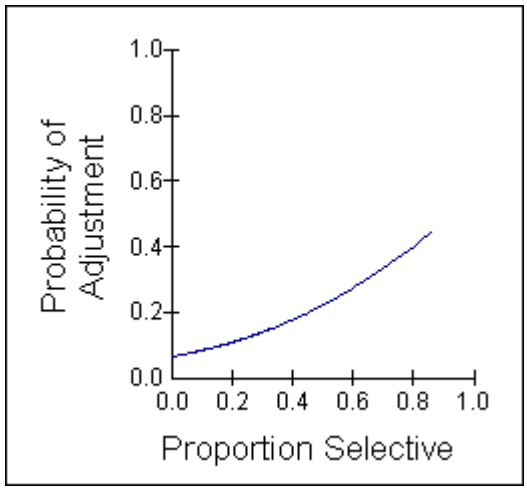
Both Long's (1997) predictive efficiency statistic and Snijders & Bosker's (1999) R^2 statistic suggested that the multilevel logistic model fitted to predict adjustment probability did not explain much variation in the data (See Appendix B). Furthermore, tests for extra-binomial variation implied under-dispersion. In other words, the assumed examiner level variance was higher than that derived from the empirical data and therefore the unconstrained estimate of examiner level variance was less than one.[5] It is quite possible that this under-dispersion occurred because of a lack of heterogeneity in the

[5] Logistic regression models assume that the observed responses are binomially distributed, that the underlying probability of an event is the same for all individuals within the population and that individuals behave independently. To effect the first of these assumptions, a condition is placed upon the level 1 variance which constrains it to equal one.

outcome measure. Indeed, there were no adjustments made to the marking of any examiners in seven of the 21 units under consideration.

Although there was no significant unit level variation in the probability of an adjustment, the one factor that did appear to affect the dependent variable was the proportion of selective candidates that an examiner had in his or her allocation. The larger this statistic, the higher the probability of an adjustment (Figure 6).

FIGURE 6 The modelled probability of an adjustment dependent upon the proportion of candidates from selective/independent centres within an allocation



Because of the potential subjectivity in the adjustment procedure, it would be tempting to suggest that those involved in the decision making process were influenced in their decisions by the characteristics of an examiner's allocation. However, despite knowing the examiner's centre range, the officers had no indication of the centre name or centre type. Moreover, this idiosyncratic relationship between centre type and marking accuracy was reproduced in *Model 1* and *Model 2* where the scope for subjectivity was even more limited.

Model 4 – The Examiner Performance Rating Model

At the end of the marking period, an examiner performance record is completed. As part of this record, senior examiners are required to assign an overall classification for each examiner according to the following instructions:

(A) Consistently excellent - The examiner marked consistently, complied with the various requirements and regulations, and did not need any special attention.
(B) Sound and reliable - The examiner will usually have complied with all the requirements and marked consistently, but might have needed some guidance at the first phase sample stage.
(C) Satisfactory: some errors but within tolerances - The examiner is one who marked consistently in the end, but needed a fair amount of help from the senior examiner and may have needed to submit an additional first phase sample or may have needed adjustment. (He/she might be inexperienced, or might have been stretched by the task, perhaps because of too large an allocation.) He/she should be retained in the expectation of improvement.
(D) Grounds for concern: re-training to be considered - The examiner caused significant difficulties during the sampling process, for example, by marking

inconsistently or misinterpreting the mark scheme, and may have been allowed to continue only after submitting an additional first phase sample. At the second phase sample stage the senior examiner remarked twenty-five or more, rather than fifteen, scripts. The examiner will normally be required to undertake some re-training.

(E) Unsatisfactory: not to be re-employed - All examiners whose work was remarked come into this category, as do those whose behaviour caused substantial difficulties. They will not normally be offered re-appointment.

(AQA, 2003)

As with the evidence assimilated to determine whether or not an examiner adjustment should be applied, the information used to assign an examiner performance rating necessarily includes elements of subjectivity. Indeed, as opposed to suggesting features of examiner background which might influence suitability to the marking task, the model fitted to predict performance rating highlighted the significance of the various sources of information used in the decision making process. The difference between assistant examiner and senior examiner mark was not among these sources, nor was the proportion of independent or selective centres in an examiner's allocation.

Size of allocation and years since appointment both proved statistically significant and are both inextricably confounded with examiner performance (Table 5). The number of scripts allocated to an examiner is based, to a certain extent, on that examiner's performance in the previous year. For every hundred extra scripts allocated the model predicted an increase of 0.1 in the performance rating[6]. Similarly, the longer an examiner has worked for AQA, the more likely that examiner marks to a high standard. For every extra year of continuous service, the predicted examiner performance rating increased by 0.046. Therefore, despite being significant in a statistical sense, in a practical sense, neither allocation size nor years since appointment impacted greatly on rating.

TABLE 5 Parameter estimates from the multilevel model, modelling the examiner performance rating

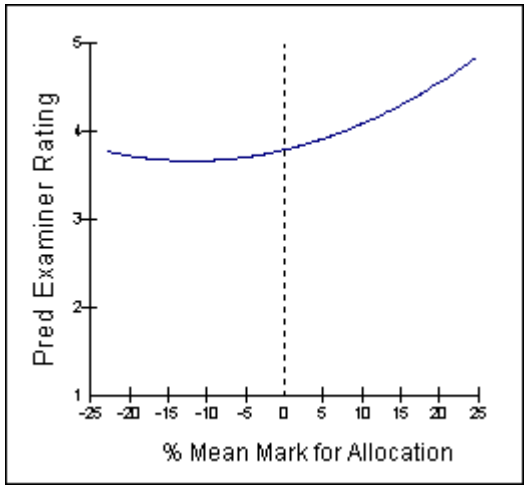
		β	Se	p
Fixed Effects	**Constant**	**3.794**	**0.087**	**0.000**
	Allocation	**0.001**	**0.000**	**0.000**
	Years since Appointment	**0.046**	**0.014**	**0.001**
	% Mean of Allocation	**0.021**	**0.006**	**0.000**
	% Mean of Allocation^2	**0.001**	**0.000**	**0.032**
Random Effects	Unit Level $\sigma^2_{Constant, Constant}$	0.033	0.024	0.162
	Examiner Level $\sigma^2_{Constant, Constant}$	**0.634**	**0.051**	**0.000**

Statistically significant effects (α=0.05) emboldened

Interestingly, the percentage mean mark of an examiner's allocation appeared to influence the examiner performance rating. The relationship described by the significant parameter estimates was quadratic (Figure 7). There was a tendency for the performance of an examiner to be perceived more favourably if the percentage mean mark for his or her allocation was higher. It seems unlikely this effect was attributable to a preference for relatively lenient marking because, during the initial model fitting exercise, the introduction of the independent variable measuring the difference between assistant examiner and senior examiner mark proved insignificant.

[6] The examiner performance rating was treated as a continuous variable with A=5, B=4, C=3, D=2 and E=1.

FIGURE 7 The relationship between percentage mean mark for allocation and predicted examiner performance rating



When senior examiners grade assistant examiners, they are not aware of any adjustment strategy, nor have they seen any allocation-wide statistics. It is only as part of the adjustment decision making process that any feedback on examiner statistics is available. At the adjustment stage, it is conceivable that a bias could be introduced if AQA officers had a predilection for work with a higher mean mark, although *Model 3* provided no evidence to support this assumption. However, at the grading stage, senior examiners do not have sufficient evidence to introduce such a systematic bias. Furthermore, as *Model 2* illustrated, the higher the grade awarded to a piece of work the less accurate the marking was liable to be.

The significance of the percentage mean mark for the allocation, which was observed irrespective of marking accuracy, suggested that the senior examiners might perceive the work of higher ability candidates more favourably no matter how closely the marking followed the mark scheme. Thus, the examiner performance ratings were higher for examiners marking allocations of a higher standard.

Questionnaire to Principal Examiners

Three out of the four models fitted as part of this study suggested that the composition of an examiner's allocation, in terms of the centre type of entry, had a significant effect upon marking accuracy:

Model 1 – The Percentage Difference Model

Compared with all other educational establishments, the marking applied to selective and independent centres appeared to be consistently more severe. For candidates awarded a grade D or higher, the marking applied to work from selective or independent centres appeared more accurate. On the other hand, for the lower ability candidates, the marking applied to work from other educational establishments appeared more accurate.

Model 2 – The Absolute Percentage Difference Model

The marking applied to scripts from selective or independent centres was slightly less accurate than that applied to scripts from other educational establishments.

Model 3 – The Adjustment Model

The one factor that did appear to affect the dependent variable was the proportion of selective candidates that an examiner had in his or her allocation. The larger this statistic, the higher the probability of an adjustment.

As assistant examiners have few details about the centres included in their allocations, the cause of the observed pattern was difficult to discern. Advice was sought from the Principal Examiners responsible for the units under consideration. A short questionnaire was drafted asking for their views on the findings (Appendix C). The covering letter dispatched with the questionnaire was left deliberately vague. Although it alluded to the fact that marking reliability had been shown to differ for selective/independent centres, it did not explain the nature of this difference. Fifteen out of the nineteen Principal Examiners contacted sent back a response.[7] Eight assumed that the marking applied to these centres was more accurate, writing comments such as:

> ' …… candidates from …… selective centres are taught better or just exam drilled more thoroughly, or taught to set work out according to a more easily recognisable structure. Examiners working with more "standard" solution layouts and approaches will find it easier to assign the corresponding marks more accurately', and
>
> 'If it is established that teachers in selective schools train their pupils to express themselves more clearly, then we might expect examiners' performance to be more accurate when marking scripts from selective schools.'

Some respondents suggested that presentation, clarity of arguments, handwriting and coherence might all affect marking reliability, implying that these qualities might be more prevalent in scripts from selective and independent centres.

Although *Model 1* provided a slightly ambiguous picture of the relationship between marking accuracy and centre type of entry, the effects identified in *Model 2* and *Model 3* suggested that the work of candidates from selective or independent centres was marked less accurately. Therefore arguments linking selective or independent education with ease of marking were not wholly supported by empirical evidence.

There were, however, six Principal Examiners who suggested marking accuracy might be compromised for candidates educated within the selective and independent sector. The respondents presented two separate arguments. The first appeared to relate to the mark scheme and its application. The second to the homogeneity of responses from candidates entered through selective and independent centres. Both arguments assumed candidates from these centre types produced a higher quality of work than that produced by candidates entered through other educational establishments. The latter additionally assumed that, within centre, the spread of marks was lower for selective and independent centres. In the current study both assumptions held true. In fact, the mean mark for candidates from selective and independent centres was almost 6% higher than that for other centre types. Examples of each argument are given below:

> ' ….. if …… candidates' work is in the upper ends of the performance spectrum there is typically quite a wide range of marks within which a mark can be allocated and therefore more scope for differences from the Principal Examiner's view',
>
> ' …… but sometimes …… [a selective or independent] centre drills candidates and work is very similar – this can lead to examiner exasperation and, instead of marking each candidate on his or her merits, they start to punish the centre', and
>
> ' …… if you have marked 200 mediocre scripts and get a good centre you fall into one of two schools – feeling heartened you give away marks because the language

[7] Although there were 21 units under consideration, papers for these units were set by only 19 principal examiners.

> is good and you make assumptions they know what they are talking about …… or the converse – so used to not seeing the points made that you miss the valid well written alternatives.'

A mark scheme is designed to provide a set of rules by which marks can be awarded and therefore, dependent upon its fitness for purpose, the mark scheme could also affect marking accuracy. In question, however, is the extent to which it could produce a systematic effect such as that seen for the selective and independent centres in this study.

In A Level examinations the mark scheme caters for all possible valid responses and generic examples are often included. These examples may work to the detriment of candidates with performances at the extremes of the mark distribution. Furthermore, levels of response marking may produce differing accuracy across the mark range if the marks available within each level are not equal. It is widely recognised that examiners are sometimes reluctant to award marks across the full mark range. With a view to easing the grade awarding problems caused by this phenomenon, several awarding bodies have instigated research in this area. It is possible that, as a by-product of this research, efforts to extend the range of marks used by all examiners may also improve marking accuracy for candidates entered through selective and independent centres. Changes to the mark scheme would effectively increase the heterogeneity of responses from selective and independent schools. Rather than viewing all pieces of work from one centre as uniform or formulaic, the examiner would be given the tools to distinguish between responses.

Thus arguments presented in response to the questionnaire, citing homogeneity of scripts and ineffective mark schemes as contributing to lower marking accuracy, support the model findings. Nevertheless there is clearly a need to determine whether the pattern observed within the sample of units under consideration is repeatable. One of the responding Principal Examiners expressed concerns with the study design asking,

> 'In the case of 2nd sampling, examiners choose the centres to send to the Principal Examiners (and Team Leaders). Has any analysis been done on the weighting of such samples re selective/independent v non-selective centres?'

Indeed any bias in the sample might influence the conclusions drawn from the model and Goldstein (1995) noted that:

> 'Although the direct modelling of clustered data is statistically efficient, it will generally be important to incorporate weightings in the analysis that reflect the sample design or, for example, patterns of non-response, so that robust population estimates can be obtained ….'

This matter is discussed in further detail in Pinot de Moira (2002). An analysis of the data used in the current study suggested that, although there was a significant difference in the sample proportion of selective and independent centres compared with the population proportion, there was no systematic bias. In other words, for some units the proportion of data points from these centres was higher than that in the population and, for some, it was lower. The sample proportion over all the units under consideration was 19.5% (approximate standard error 0.5%) and the population proportion was 18.0%.

One further factor which may have contributed to the apparent lower levels of accuracy in marking of work from selective and independent centres was the inclusion of a small minority of data points from the post-award marking checks (see footnote 2, page 8). Some of these checks are initiated on the

basis of the difference between centre estimates and final grade awarded. If centres from the selective and independent sector were more likely to over-estimate the performance of their candidates, they may also have a higher probability of inclusion in a post-award check and, therefore, have a greater chance of mark changes to the work of their candidates. In fact, recent work completed on a sample of the new Curriculum 2000 A Level examinations suggested that 'the extent to which teachers' mean estimates exceeded mean awarded grades was significantly smaller within Independent schools and both kinds of Selective Secondary schools' (Dhillon, 2003). It is unlikely therefore, that the inclusion of post-award data points affected statistical significance of the centre type effects observed in *Model 1*, *Model 2* and *Model 3*.

CONCLUSIONS

Current research into the background of examiners and the marks they give was inspired by recommendations made in the independent panel report on maintaining GCE A Level standards (Baker et al., January 2002). Despite being designed as a pilot, the results provide a valuable insight into the external factors which affect marking reliability. The key finding is that the composition of an examiner's allocation has far more influence on accuracy than easily measurable features of an examiner's background. To a certain extent, the pilot therefore fails in its aim to identify factors which might allow awarding bodies to predict those assistant examiners who are likely to be most efficient or, on the other hand, those who are likely to require additional training or monitoring.

The models produced some evidence to support the argument that the work of better candidates was harder to mark, as was the work of candidates from independent and selective establishments. Less able candidates from these centre types appeared to be at a particular disadvantage compared with their counterparts educated elsewhere.

Some of the independent variables which were shown to impact significantly upon marking reliability were confounded with the outcome measures. For example, in general, only examiners who mark to a high standard continue to be invited to assess the work of candidates, hence the significance of years since appointment in two of the models. Implicit in awarding body procedures is an informal decision making process which refines the pool of examiners. The officers involved in the delegation of work use many of the variables included in these analyses to determine: whether to adjust the marking of an examiner; whether to retain an examiner or how large the examiner's allocation should be. The senior examiners responsible for the assignment of examiner performance ratings also have some of this information available.

In all of the models, with the exception of *Model 4*, there remained significant unit level variation in the data after the final model had been fitted. Marking accuracy differed between subject areas and whilst, for many units included in the study there was a good degree of agreement between the assistant examiner and the senior examiner mark, in some subject areas the discrepancies were larger. With unit - as opposed to examiner - characteristics in mind, therefore, the pilot study succeeds in identifying areas where extra training and monitoring resources might be concentrated. The lack of unit level variation in *Model 4* lends further weight to the need for additional training. It suggests that, no matter what the overall quality of marking within a unit, senior examiners appear able to identify assistant examiners with the qualities required of each rating. In other words, the need to identify a hierarchy of examiners for future monitoring purposes, may drive senior examiners to award performance ratings above that which might be suggested by the rating definitions. The senior examiners effectively seem to rank order examiners within their team rather than to rate them. For some units, this might suggest an unjustified level of satisfaction with the marking.

RECOMMENDATIONS

As a pilot to ascertain the feasibility of a methodology to identify examiner characteristics influencing marking reliability, the models fitted in this study fail in their aim. Nevertheless, the findings provide a rich vein of information regarding features which may affect the accuracy of marking. Before operational recommendations can be made, however, further research questions need answering. There is a need to ascertain whether the results presented herein are repeatable. The analyses were performed on a relatively small sample of data. The data were taken from the non-random, second phase sample rather than drawing a random mark-remark sample. Only 21 units were included in the analyses and these were all A2 units. There were no examples of scripts examined at any other level. Particularly for *Model 3*, the extent to which the independent variables explained variation in the dependent variable was minimal and, for all models, the residual plots highlighted a number of outliers (Appendix B).

The limitations of the current study suggest improvements for future studies. A decision needs to be taken as to whether the initial central aim of this study – to identify links between examiner backgrounds and the marks they give – is still of prime interest or whether further investigation of the features of examiner's allocations is now the priority. If the initial aim remains the focus, it is recommended that:

- More detailed information about examiner backgrounds should be collected, possibly by means of a questionnaire or by use of an off-the-shelf or bespoke personality measurement tool.
- Further consideration should be given to future use of the analysis. If it is proposed that the derived models are used as part of the recruitment process, what dependent variable provides the best operational measure of marking accuracy and which independent examiner level variables will be available at the time of recruitment? The same type of questions need to be asked if the derived models are to be used to allocate training and monitoring resources.

On the other hand, if the findings from the current study have shifted attention to the composition of allocations, it is recommended that:

- The study should be repeated using a more robust data set and taking examples of work from, at least, GCSE, AS and A Level examinations and across a full range of subject areas.
- The theoretical validity of using each of the available independent variables should be addressed, particularly in view of the fact that future studies could include past performance ratings.
- Further investigation should be made into the possibility of non-linear relationships between the independent and dependent variables and into the existence of interaction effects.
- After the study has been repeated, conclusions should be drawn as to which of the fitted models provides the greatest understanding of the relationship between the features of an examiners allocation and marking reliability.
- Some consideration should be given to how the findings from the modelling exercise could be used to create operational rules designed to maximise marking accuracy.

As a by-product of the data analysis and questionnaire responses, further research questions outside the remit of the current study have arisen. With a view to adopting a more pro-active research stance, it is suggested that the following areas of study are considered by QCA alongside the recommendations for continued research made above:

- An investigation into how the quality of written communication affects marking accuracy (arises from questionnaire responses).
- An experiment to determine whether examiners perceive quality of marking to be higher if the quality of the script is higher (arises from *Model 4* findings).

REFERENCES

AQA. (2003). Senior Examiner's Handbook for Standardisation Meetings and Supervision of Examiners.

Archer, J., & McCarthy, B. (1988). Personal biases in student assessment. *Education Research, 30*(2), 142-145.

Baird, J. (1998). What's in a name? Experiments with blind marking in A-level examinations. *Educational Research, 40*(2), pp 191 - 202.

Baird, J., Greatorex, J., & Bell, J. F. (2002). What makes marking reliable? Experiments with UK examinations. *RC Papers, RC/191*.

Baird, J., & Pinot de Moira, A. (1997). Marking reliability in Summer 1996 A Level Business Studies. *RAC Paper, RAC/760*.

Baker, E., McGraw, B., & Lord Sutherland of Houndwood. (January 2002). Maintaining GCE A Level standards: The findings of an independent panel of experts: QCA.

Bell, J. (2000). Statistical detection of lingering doubt examiners.

Branthwaite, A., Trueman, M., & Berrisford, T. (1981). Unreliability of marking: further evidence and a possible explanation. *Educational Review, 33*(1), 41-46.

Dhillon, D. (2003). Teachers' estimates of candidates' grades: Curriculum 2000 advanced level qualifications. *RC Papers, RC/ In Press*.

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

Goldstein, H. (1995). *Multilevel statistical models*. London: Arnold.

Hartog, P., & Rhodes, E. C. (1935). *The Marks of Examiners*: In Black, E. L., (1962) The Marking of G.C.E. Scripts. *British Journal of Educational Studies, 11*, 61-71.

Hughes, D. C., Keeling, B., & Tuck, B. F. (1980). Essay marking and the context problem. *Educational Research, 22*(2), pp. 147-148.

Long, J. S. (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks CA: Sage Publications.

Lunz, M. E., & O'Neill, T. R. (1997, March 24-28). *A longitudinal study of judge leniency and consistency.* Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, Illinois.

Meyer, L. (2000). The ones that got away – Development of a safety net to catch lingering doubt examiners. *RC Paper, RC/50*.

Murphy, R. J. L. (1978). Reliability of marking in eight GCE examinations. *British Journal Of Educational Psychology, 48*, pp. 196-200.

Pal, S. K. (1986). Examiners' Efficiency and the Personality Correlates. *Indian Educational Review, 21*(1), pp. 158-163.

Pinot de Moira, A. (2002). Statistical robustness in comparability studies: The choice of model and data selection. *RC Papers, RC/172*.

Pinot de Moira, A., Massey, C., Baird, J., & Morrissy, M. (2002). Marking consistency over time. *Research in Education, 67*, 79-87.

Powers, D., & Kubota, M. (1998a). *Qualifying Essay Readers for an Online Scoring Network (OSN)* (RR-98-20). Princeton, New Jersey: Educational Testing Service.

Powers, D., & Kubota, M. (1998b). *Qualifying Readers for the Online Scoring Network: Scoring Argument Essays* (RR-98-28). Princeton, New Jersey: Educational Testing Service.

QCA, ACCAC, & CCEA. (2002). *GCSE, GCSE in vocational subjects, GCE, VCE and GNVQ Code of Practice 2002/03*.

Rasbash, J., Brown, W., Goldstein, H., Yang, M., Plewis, I., Healy, M., Woodhouse, G., Draper, D., Langford, I., & Lewis, T. (2000). *A user's guide to MLwiN* (2.1 ed.): Multilevel Models Project.

Scottish Qualifications Authority. (n.d.). *Scotland's pupils need a good ticking off: Application for marking duties*. Retrieved on 28th April 2003 from www.sqa.org.uk/files_ccc/

Markers%20_app_form_02.pdf.

Shaw, S. (2002). The effect of training and standardisation on rater judgement and inter-rater reliability. *Research Notes: University of Cambridge Local Examinations Syndicate, May*, 13-17.

Snijders, T. A. B., & Bosker, R. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*: Sage Publications.

Spear, M. (1996). The influence of halo effects upon teachers' assessments of written work. *Research in Education, 56*, pp. 85 - 87.

Spear, M. (1997). The influence of contrast effects upon teachers' marks. *Educational Research, 39*(2), 229-233.

The Quality Assurance Agency for Higher Education. (January 2000). *Code of practice for the assurance of academic quality and standards in higher education. Section 4: External examining*. Retrieved on 24th April 2003 from www.qaa.ac.uk/public/cop/copee/ COP_external.pdf.

Whetton, C., & Newton, P. (2002, 1-6 September 2002). *An evaluation of on-line marking.* Paper presented at the 28th International Association for Educational Assessment Conference, Hong Kong.

APPENDIX A Variable Formulation

Available to Model	Variable Label	Variable Description	Coding
1,2,3,4	Unit	A2 unit code	Categorical variable, 1 CPT4 2 CPT5 3 GGA4 4 GGA5 5 GGB4 6 GGB5 7 ICT4 8 ICT5 9 LTA4 10 LTA6 11 MAP1 12 MAP2 13 MAP3 14 MAP6 15 MBP2 16 MBP5 17 MBP6 18 MBP7 19 PA04 20 PA10 21 PHB5
1,2,3,4	Examnum	Examiner number	Categorical variable, personal numbers as allocated by AQA
1,2	Centnum	Centre number	Categorical variable, National Centre Numbers
1,2	Candnum	Candidate number	Categorical variable, as allocated by the centre
1,2	Centtype	Centre type	Categorical variable, 1 Secondary Comprehensive or Middle Community (Voluntary Aided) 2 Secondary Selective (Voluntary Aided) 3 Secondary Modern Controlled (Voluntary Aided) 4 Secondary Comprehensive or Middle (Foundation) 5 Secondary Selective (Foundation) 6 Secondary Modern (Foundation) 7 Independent 8 FE Establishment 9 Sixth Form College 10 Tertiary College 11 Other (including private candidates) 12 Overseas
1,2	Centgrp	Centre type (grouped)	Categorical variable, 1 Schools 2 Selective or Independent 3 Further Education (FE)
1,2	Sexcand	Candidate sex	Binary categorical variable, 0 Male 1 Female
1,2	Candage	Candidate age (in years)	Continuous variable, centred around 18 giving values in the range -3 to 65
1,2	Candageg	Candidate age (grouped)	Categorical variable, 1 <18 2 18 3 >18
1,2	Gradeawd	Grade awarded to the candidate for the unit	Added as both a continuous and categorical variable, 1 U 2 E 3 D 4 C 5 B 6 A

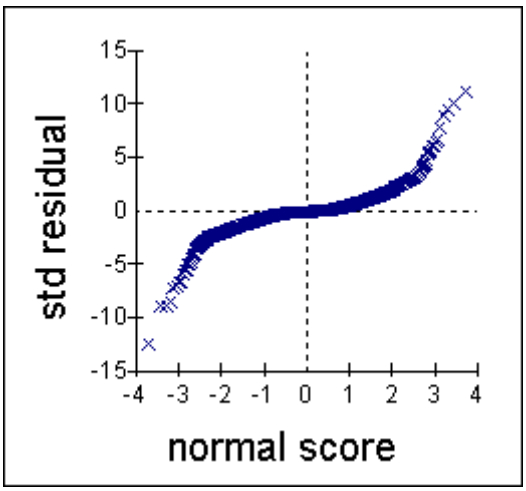
Available to Model	Variable Label	Variable Description	Coding
1,2	Gradeest	Grade estimated for the candidate in the unit	Added as both a continuous and categorical variable, 1 U 2 E 3 D 4 C 5 B 6 A
-	Semark	Senior examiner mark	Used only to formulate the dependent variable
-	Rawmark	Raw assistant examiner mark	Used only to formulate the dependent variable
1,2,3,4	Sexexam	Examiner sex	Binary categorical variable, 0 Male 1 Female
1,2,3,4	Qual	Examiner qualification	Categorical variable, 1 BEd 2 BSc/BA 3 PGCE 4 MSc/MA/MEd 5 DPhil/PhD 6 Other/Unknown
1,2,3,4	Examjob	Examiner job (13 categories)	Categorical variable, 1 Head/Deputy 2 Head of Department (HOD) 3 Teacher 4 Assistant Teacher 5 Assistant Head of Department (Asst HOD) 6 Part Time 7 Retired 8 Self-Employed 9 Supply 10 Lecturer 11 Director 12 Co-Ordinator 13 Other/Unknown
1,2,3,4	Jobcat	Examiner job (9 categories)	Categorical variable, 1 Head/Deputy 2 Asst/HOD 3 Asst/Teacher 4 Part Time 5 Retired 6 Self-Employed 7 Lecturer 8 Co-Ordinator 9 Other/Unknown
1,2,3,4	Yearsapp	Years since appointment to AQA (or AEB/NEAB)	Continuous variable, centred around 3 giving values in the range -3 to 7
-	Examrnk	Examiner rank	Categorical variable (not used as an independent variable), 1 Examiner 2 Principal Examiner (PE) 3 Team Leader
3	Adj	Indicator denoting whether an adjustment was made to the examiner's marking	Binary categorical variable 0 No 1 Yes
1,2,3,4	Alloc	Size of examiner's allocation	Continuous variable, centred around an allocation of 350 scripts giving values in the range -306 to 528

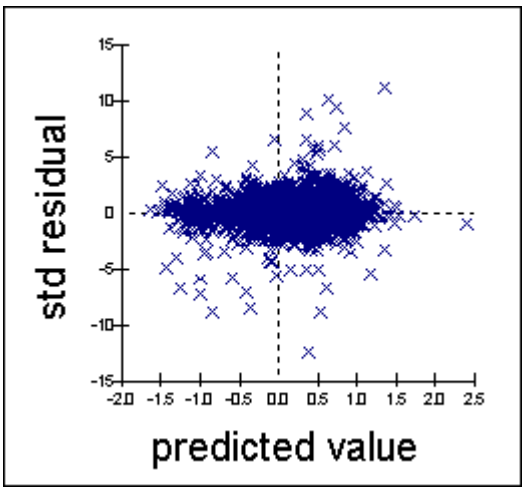
Available to Model	Variable Label	Variable Description	Coding
1,2	Candcent	For each centre, the number of candidates entered for the unit	Continuous variable, centred around a centre entry of 30 candidates giving values in the range -29 to 220
4	Examgrad	Examiner performance grade or rating	Continuous variable, 1 E 2 D 3 C 4 B 5 A
1,3,4	Markdiff	the percentage difference between assistant examiner mark and senior examiner mark	Continuous variable
2,3,4	Absdiff	the absolute percentage difference between assistant examiner mark and senior examiner mark	Continuous variable
1,2	Meangcse	Candidate mean GCSE result	Centred continuous variable. Each GCSE result is coded such that A*=8, A=7 …… U=0. A mean of all GCSE results is calculated. This statistic is centred around a mean GCSE result of a grade C giving values in the range -5 to 3.
1,2	Nogcse	Number of GCSE examinations contributing to a candidate's mean GCSE result	Centred continuous variable, centred around 9 giving values in the range -8 to 5.
1,2,3,4	Maxmark	Maximum mark for the unit	Continuous variable
1,2,3,4	Percmean	Mean of an examiners allocation, expressed as a percentage of the maximum mark for the unit	Continuous variable, centred around 50% potentially giving values in the range -50% to 50%
3,4	Propsele	Proportion of candidates marked by an examiner who are entered by selective centres	Continuous variable, centred around the mean of 0.1779 giving values in the range -0.1779 to 0.6888.
3,4	Propfe	Proportion of candidates marked by an examiner who are entered by further education centres	Continuous variable, centred around the mean of 0.3077 giving values in the range -0.3077 to 0.6923.
3,4	Propsex	Proportion of candidates marked by an examiner who are female	Continuous variable, centred around the mean of 0.3878 giving values in the range -0.3878 to 0.4581.
3,4	Under18	Proportion of candidates marked by an examiner who are under 18	Continuous variable, centred around the mean of 0.0799 giving values in the range -0.0799 to 0.8259.
3,4	Age18	Proportion of candidates marked by an examiner who are 18	Continuous variable, centred around the mean of 0.7439 giving values in the range -0.7439 to 0.1680.
3,4	Over18	Proportion of candidates marked by an examiner who are over 18	Continuous variable, centred around the mean of 0.0882 giving values in the range -0.0882 to 0.3576.

APPENDIX B Residual Plots & Model Diagnostics

Model 1

Level 1 Residuals – Candidate Level

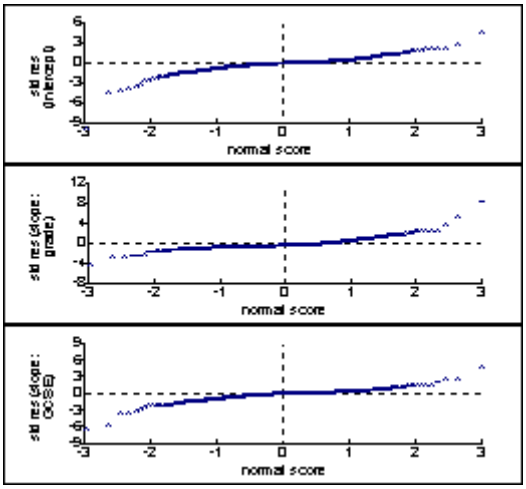


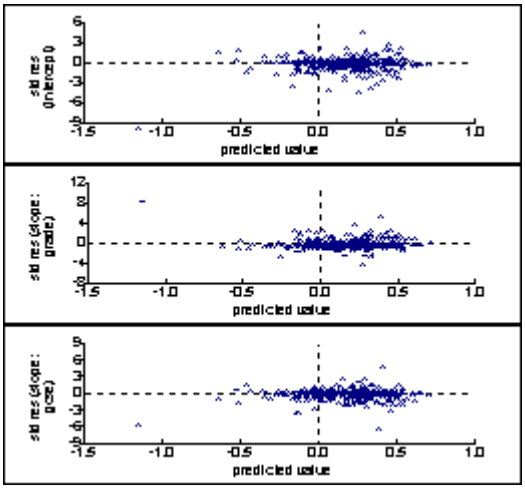


Level 2 Residuals – Centre Level

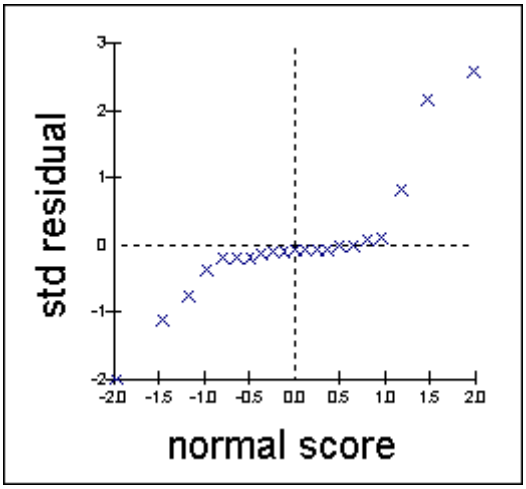
MLwiN software crashed in all attempts to calculate residuals at this level.

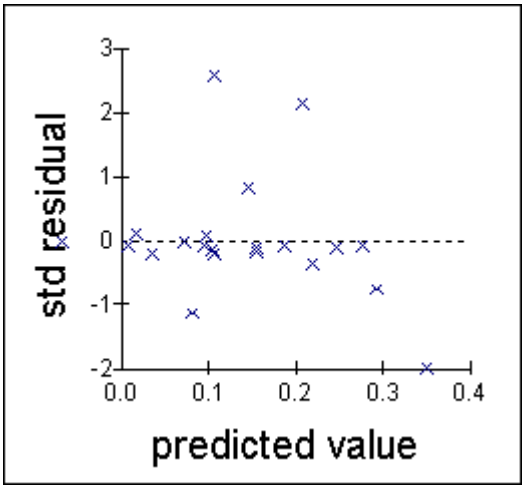
Level 3 Residuals – Examiner Level





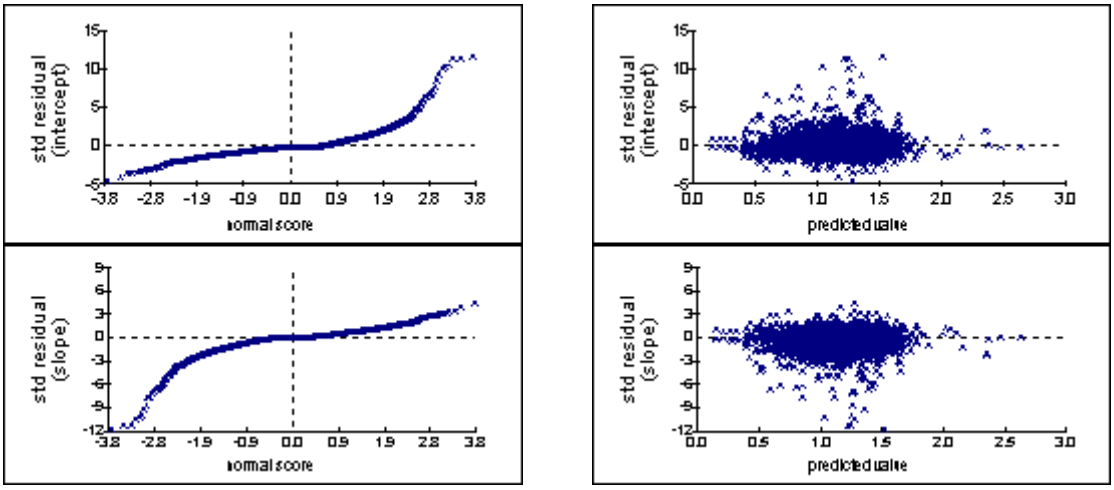
Level 4 Residuals – Unit Level



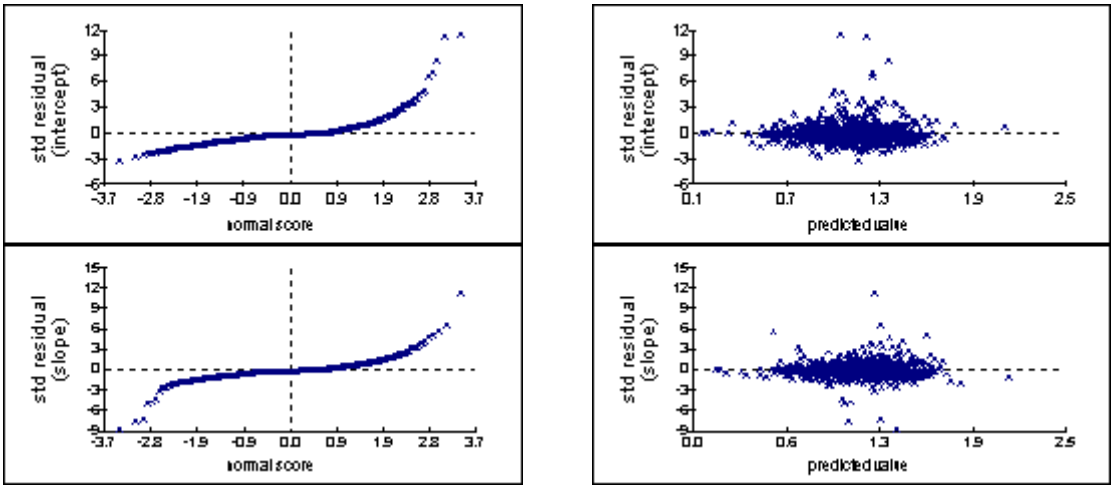


Model 2

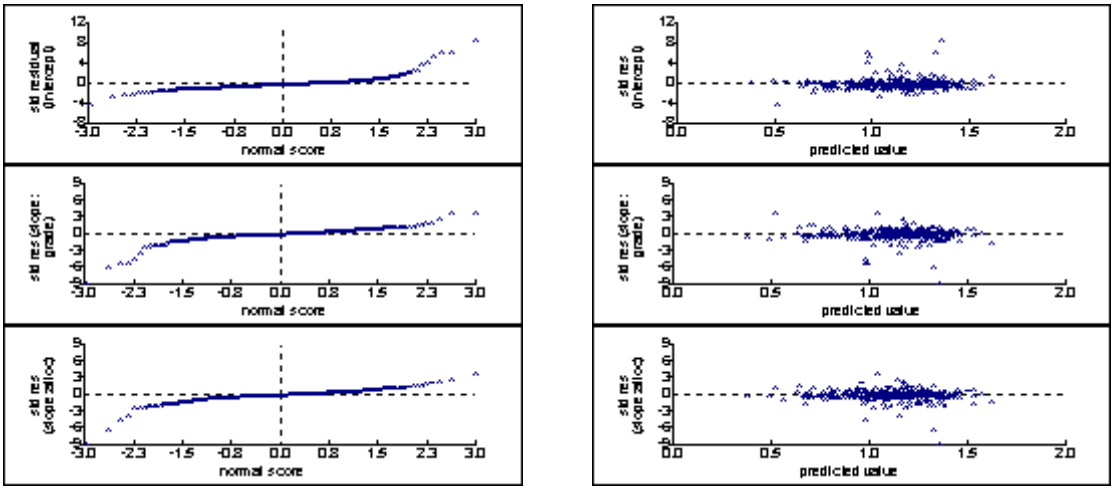
Level 1 Residuals – Candidate Level



Level 2 Residuals – Centre Level

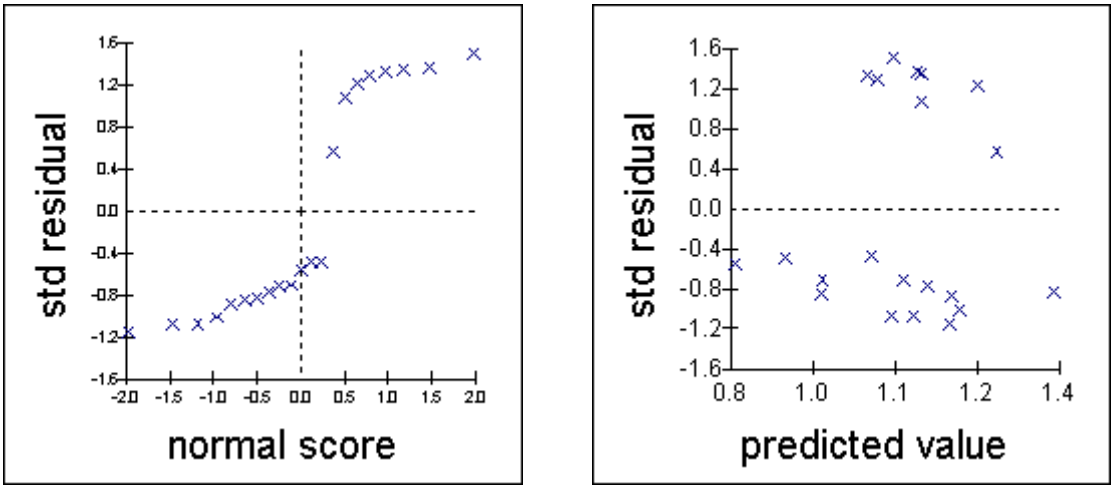


Level 3 Residuals – Examiner Level



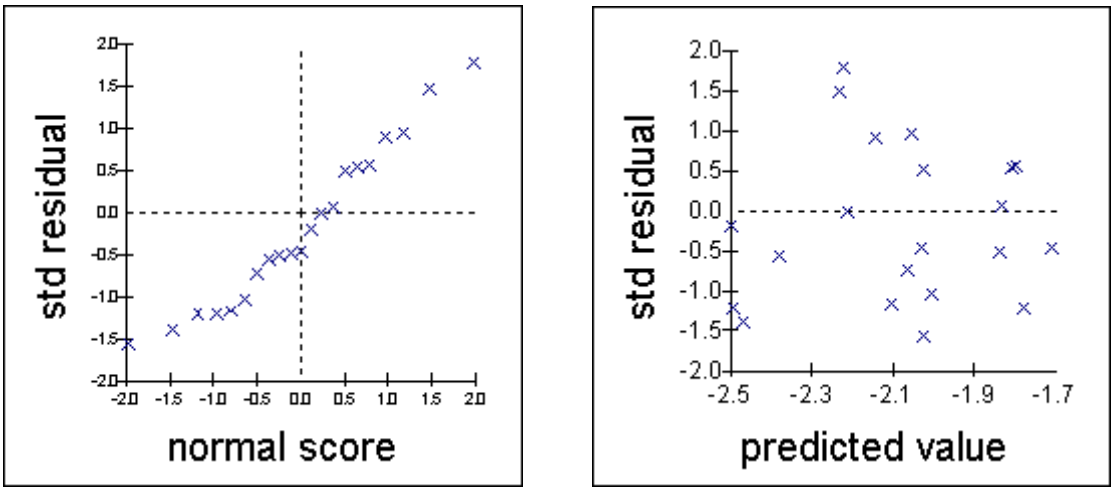
Model 2

Level 4 Residuals – Unit Level



Model 3

Level 2 Residuals – Unit Level

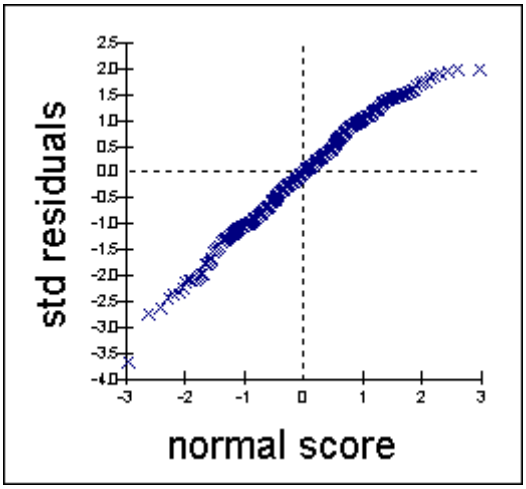


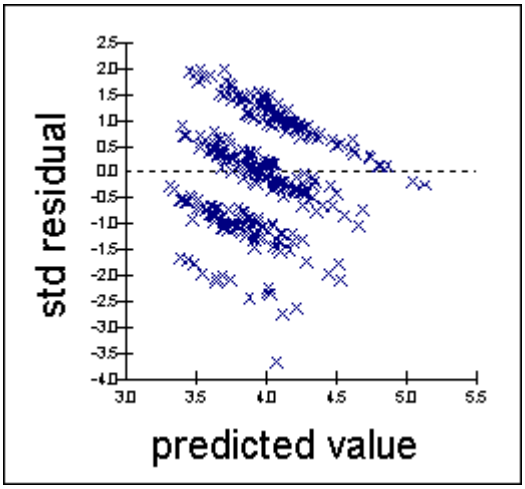
Model Statistics

Long's Predictive Efficiency	0.000
Snijders & Bosker's R^2	0.045
Extra-Binomial Variation: Parameter Estimate (Standard Error)	0.874 (0.067)

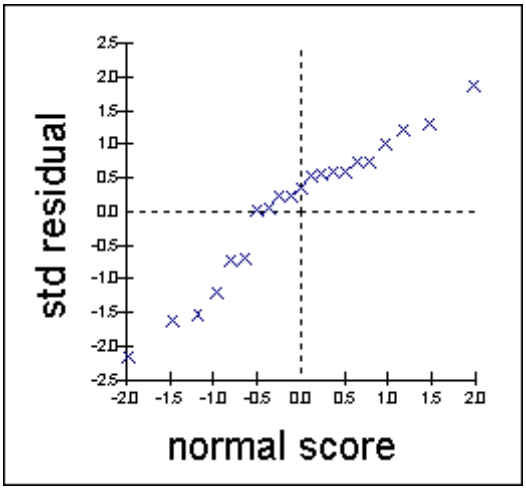
Model 4

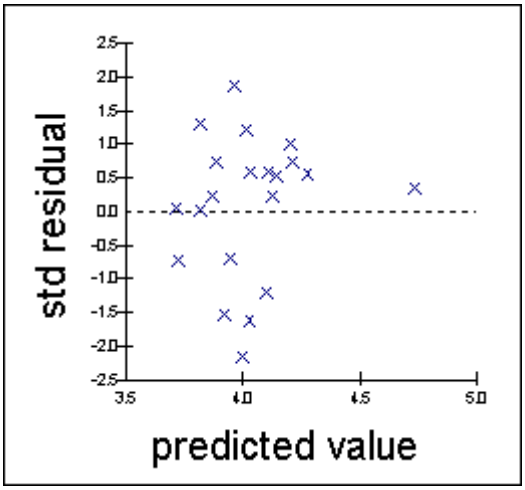
Level 1 Residuals – Examiner Level





Level 2 Residuals – Unit Level





APPENDIX C Questionnaire to Principal Examiners

<<Principal Examiner Name>>
<<Address1>>
<<Address2>>
<<Address3>>
<<Address4>>
<<Postcode>>

14 April 2003

Dear <<Principal Examiner Name>>

The AQA Research & Statistics Department has an ongoing programme of research into the reliability of marking. The work we have completed in the past has been used to improve operational procedures, feedback to examiners and design of marking schemes.

As part of our current research, we have noticed that the marking accuracy of Assistant Examiners appears to be influenced by the centre type through which a candidate is entered. In particular, marking accuracy is different for candidates from selective and independent centres. We are struggling to find a reason for the pattern that we have observed and have decided to turn to you, in your capacity as Principal Examiner, to ask whether you can shed any light on our findings.

We would be grateful if you could spare a few moments to answer the two-part question posed on the sheet enclosed with this letter. Your insight will be used to help with the interpretation of current findings. To avoid compromising the independence of your comments, we will delay providing exact details of the nature of the study until the research is complete.

A postage paid envelope is included for the return of your comments which we would be happy to receive by the **2nd May 2003**. Thank you in advance for your help.

Yours sincerely,

Anne Pinot de Moira
Senior Research Officer

Enc..

How, and why, do you think that marking accuracy of Assistant Examiners might be influenced by the fact that the candidate whose work they are marking is entered from a selective or independent centre?

How? .....

.....

.....

.....

.....

.....

.....

Why? .....

.....

.....

.....

.....

.....

Please return in the enclosed pre-paid envelope
Thank you

xxxxx