STATISTICAL ROBUSTNESS IN COMPARABILITY STUDIES
The Choice of Model and Data Selection

Anne Pinot de Moira

1 INTRODUCTION

Over the past few years, logistic multilevel models have been used to model the probability of candidates exceeding a given grade boundary in a given subject dependent upon the awarding body of entry. Questions remain over the validity and robustness of such an approach, not simply from a modelling point of view but also from a data selection point of view. Many of these questions were discussed at a seminar addressing the methodologies applied in recent comparability studies (Fowles, 2000) but, while differing views were freely expressed, there was little opportunity to consider any supporting statistical evidence. This report collects together research from the areas of multilevel modelling, non-linear modelling and linear modelling to provide a background to the techniques currently used in inter-awarding body comparability studies. Little is made of whether a statistical approach to the comparison of grading standards is, in fact, appropriate because this matter was discussed in detail by Baird & Jones (1998) and by Baird, Cresswell et al (1999). Baird & Jones (1998) concluded that,

> *"Examination standards cannot be measured because they cannot be extricated from the characteristics of the examinations themselves or from the characteristics of the people taking the examinations. Value judgements about comparisons between examination standards will still have to be made in the design and interpretation of statistical analyses of examination comparability."*

Nevertheless Baird, Cresswell et al (1999) argued that,

> *"[Since they will always] be required to defend their maintenance of [standards] from a range of conflicting perspectives …… it is clearly essential for the boards to be open about the problematic nature of examination standards and the processes by which they are determined."*

With this warning in mind, issues surrounding continuing good practice in, and improvements to, the statistical comparison of grading standards are discussed.

2 UNRESOLVED ISSUES

In any comparability study, the limiting factors include time, cost and available data. Therefore, the search for an ideal statistical technique by which to assess the comparability of grading standards between awarding bodies must combine statistical validity with pragmatism. This report considers some of the unresolved issues surrounding data and model selection by referring to the data available from the GCSE English comparability study (Pinot de Moira, 2000).

2.1 The Model

2.1.1 The Dependent Variable

To be of practical value, the statistical output from a comparability study must provide a clear indication of any necessary remedial action. The choice of model plainly affects the ease with which recommendations can be implemented. To compare the grading standards between awarding bodies the outcome or dependent variable must be the grade achieved in the subject area of interest. Fielding (1999) discussed the way in which such a dependent variable should be formulated and some of his arguments are pertinent within the context of inter-board comparability studies.

Grade is an ordered categorical variable but is often assumed to be on an equal interval scale and therefore treated as continuous. This assumption might appeal because it allows candidate grade to be modelled using linear regression. The principals of linear regression are familiar and, in a statistical sense, the interpretation of parameter estimates is straightforward. The model can be used to predict the effect a one unit increase in the independent variable will have on the dependent variable. If the independent variable is a contrast between, for example, AQA and other awarding bodies then the model may predict that candidates entered through the former are awarded an average of a seventh of a grade less than those entered through the latter. Such a finding does not, however, provide specific information about how to redress the inequity between awarding bodies. It suggests that the inequity exists uniformly across the whole grade range. In award meetings, where grading standards are determined, the committee is required to make judgements about the placement of several specific grade boundaries. To an extent these decisions are independent and based upon evidence pertinent to the boundary under consideration. In this way, the awarding committee is able to account for skewed performances across the mark range. Therefore, on the basis of comparability study findings, to suggest to that **all** boundaries should become more severe (or more lenient) is to ignore the possibility that differences between awarding bodies may be grade specific. Furthermore, to implement the findings from a linear model would be to accept that the grade outcome data are continuous and that the difference in candidate achievement between any adjacent pair of grades was identical.

Since grade is an ordered categorical variable, the most natural way to model grade outcome would be to use a non-linear ordered multinomial regression. For each grade, therefore, it would be possible to estimate the relative grading standard of each awarding body in terms of the cumulative proportion of candidates at each grade. Such output could be used to inform the decisions made in future award meetings.

Even though the ordered multinomial model provides the information needed to effect remedial action, the interpretation of output is complex as predictions are based upon interactions between awarding body and grade. Furthermore, convergence of the iterative routines is dependent upon a sufficient number of observations in each outcome category. In the A Level and GCSE qualifications, grade distributions are never uniform[1] and so convergence of ordered multinomial models in the context of comparability studies is always potentially problematic. To aid convergence, collapsing the outcome variable across grades to create larger categories would still produce models which could inform future award meetings. At the ultimate extreme, candidates could be classified according to whether or not they exceed a given grade threshold and a logistic model fitted to the data. Using a simple example, Bell (1999) demonstrated that such a binary response model provided an appropriate description of the data. Indeed, this is the approach that has been taken in recent comparability studies. Although it has meant that each of the judgmental boundaries must be considered in a separate model, no problems have been encountered with convergence and, by design, the models are free from complex interactions between awarding body and grade.

While it is clear that both the ordered multinomial and logistic models provide a suitable tool with which to evaluate and address grading standard differences between awarding bodies, for transparency and, presently, for ease of construction, the logistic model is marginal favourite. Appendix A illustrates the issue of dependent variable choice by reference to examples using data from the GCSE English comparability study.

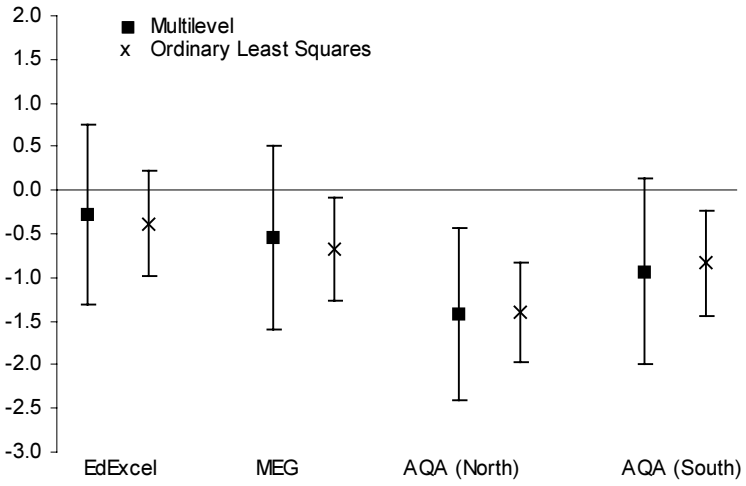
[1] In fact, in higher tier GCSE examinations where there is an allowed grade E, the grade distributions are designed such they are not uniform, lending weight to the argument that treating grade outcome as a continuous variable is far from ideal.

2.1.2 Modelling Hierarchies

The candidate-level data collected for comparability studies is hierarchically structured because the candidates are nested within centre. The extent to which a centre influences candidate outcome is therefore of interest to determine whether this hierarchy should be modelled. Using an ordinary least squares technique to make inter-awarding body comparisons requires the assumption the individual observations are independent. The higher the intra-centre correlation the greater the violation of this assumption and the less appropriate a non-hierarchical approach. Where intra-centre correlation exists, failure to model the inherent hierarchy leads to an underestimation of the standard errors associated with the parameter estimates (Goldstein, 1995; Hox, 1995). Such underestimation leads to an inflation of type 1 errors where the null hypothesis of no difference between awarding bodies is erroneously rejected (Kreft & de Leeuw, 1998). Barcikowski (1981) shows that even where intra-class correlation[2] is lower than 0.05 the effective critical value in the test of the null hypothesis can be greatly increased, especially if the number of level 1 units per level 2 unit is high.

In the GCSE English comparability study, the intra-centre correlation for the data used to create the higher tier grade C model was 0.32 and the average number of candidates per centre was 62. Both these statistics suggest that an ordinary least squares approach to the analysis could unrealistically exaggerate the differences between awarding bodies. Indeed, when the non-hierarchical and hierarchical approaches are compared, the confidence surrounding the estimates of awarding body effect in the former model are somewhat narrower than the confidence intervals associated with the latter model (Figure 1).

FIGURE 1 Awarding body parameter estimates (and associated 95% confidence intervals) from the GCSE English higher tier grade C data. A multilevel approach compared with an ordinary least squares approach



Despite worries over the lack of independence between level one units and over the increased probability of type 1 errors, there are still some opponents to the use of multilevel modelling in the analysis of examination outcomes. Criticisms levelled at the procedure are described in detail in Annex C of Fitz-Gibbon (1997) and the issues are also discussed in Kreft (1996).

---

[2] Intra-class correlation is the generic term for the degree of association between level 1 units within a level 2 unit. For a two level logistic random intercept model with an intercept variance of σ_U^2, the intra-class correlation is

$$\rho = \frac{\sigma_U^2}{\sigma_U^2 + \pi^2/3}$$ (Snijders & Bosker, 1999)

Multilevel models regard the sample of centres as a random selection from the population of all centres. This allows generalised inferences about the variation between centres (Goldstein, 1995). In particular, for centres with few candidates, the model borrows strength from other centres in the sample. This has an important consequence when making centre level inferences. For many centres, data are available for the full cohort of candidates. Where this is the case, and the centre is small relative to others, then the predictions for that centre may be speciously drawn towards the norm. Therefore a small centre of candidates with special needs may be predicted better results than achieved. This becomes a particular issue when considering centre level residuals to assess, for example, value-added. In league tables, the small special needs centre would then be identified as a centre where insufficient value is added.

For residual analyses, therefore, there is plainly some question over whether multilevel models provide the ideal tool from which to draw centre level conclusions. However, the comparability studies are concerned with making generalised inferences about the grading standards applied by each awarding body across the whole entry. For every small special needs centre, there will be a small independent centre. The estimates for each of these centres will be pulled towards the expected value for all centres to provide overall population estimates. It is these population estimates which are needed to provide information about the grading standards between awarding bodies. As such, it seems important that, for any analysis where population level conclusions are to be drawn, the degree of inter-dependence between candidates within centres is correctly modelled otherwise the tests comparing the grading standards between awarding bodies will not be suitably conservative.

2.2 The data

2.2.1 All Grades, All Grades Within Tier or Adjacent Grades

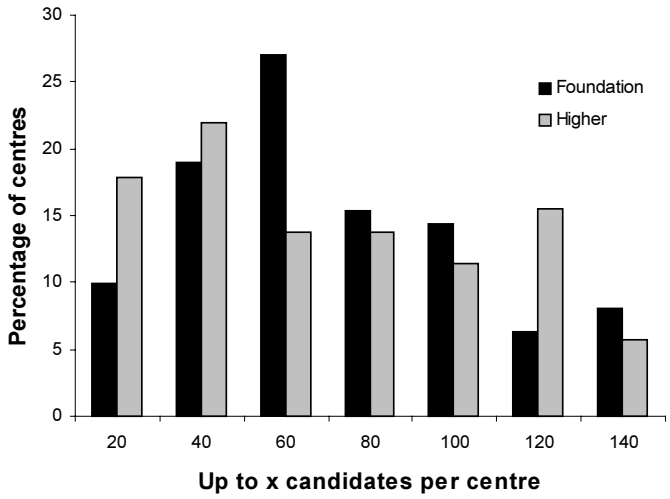
Logistic regression models assume that the observed responses are binomially distributed, that the underlying probability of an event is the same for all individuals within the population and that individuals behave independently. To effect the first of these assumptions, a condition is placed upon the level 1 variance which constrains it to equal one. A departure from this assumption might imply the existence of extra-binomial variation. The extent of the departure can be tested by relaxing the constraint and comparing the computed level 1 variance with the assumed variance of one. Extra-binomial variation is further discussed in texts including Goldstein (1995), Ramsay & Schafer (1997) & Wright (1997).

Such tests were performed in the analysis of the GCSE English data. For most of the reported models, the results of the test suggested under-dispersion in the sample data. In other words, the assumed candidate level variance was higher than that derived from the sample and therefore the unconstrained estimate of level 1 variance was less than one. The model of the sample data suggested that the candidates were more homogeneous than would be expected given the assumption that observed responses are binomially distributed. There are two possible reasons why this may have occurred.

Firstly, it is possible that, between centres, some of the variation remained unmodelled. The implication being that the model was mis-specified and that one or more important explanatory variables were omitted or a level of the hierarchy was overlooked. If this were the case then, although the parameter estimates derived from the model would not be seriously biased, the standard errors associated with these parameter estimates might be over-estimated. In the context of comparability studies, where the grading standards applied by awarding bodies is of primary interest, under-dispersion potentially increases the chance of Type II errors. The null hypothesis is more likely to be accepted and the conclusion that there is no difference between awarding bodies endorsed.

There is some evidence to suggest that spurious conclusions are sometimes drawn when extra-binomial variation is identified in models built from sparse datasets. In his paper, Wright (1997) describes sparse datasets as those with few level 1 units per level 2 unit and warns that the apparent existence of extra-binomial variation does not necessarily imply model mis-specification. Figure 2 describes the dataset used in the GCSE English comparability study. The percentage of centres which were represented by fewer than 20 candidates was low. However, there were still several centres for whom the available candidate level information was redundant in a statistical sense. Moreover, applying the definition used by Wright (1997), the distribution of candidates per centre was not uniform. Implicitly, Wright (1997) proposes the uniform dataset as the strongest safeguard against spuriously concluding model mis-specification when extra-binomial variation is identified. The data used for the GCSE English comparability study were certainly not uniform. It is possible therefore that the data structure may have contributed to the under-dispersion observed in the unconstrained model.

FIGURE 2 The distribution of candidates per centre in the dataset used for the GCSE English comparability study



The second possible explanation for the under-dispersion seen in the models derived from the GCSE English data was a bona fide lack of heterogeneity in the sample data. In other words, the observed probability of an outcome was very small or very large. In a discussion of the implications of modelling binary outcomes using multilevel techniques, Goldstein (1995) observes:

> *"….. when the average observed probability is very small (or very large), …… we will often find that where the response is binary, there will be many level 2 units where the responses are all zero [(or all one)]. In such a case, convergence often may not be possible and, even where estimates are obtained, in general they will not be unbiased. This problem can be avoided by having sufficient number of level 2 units where there is adequate response heterogeneity ….."*

This finding is certainly pertinent to the inter-awarding body comparability studies since the average observed probability of foundation tier candidates exceeding the grade F threshold, or higher tier candidates exceeding the grade C threshold, is naturally relatively high. As the data are clustered within centres, the chances of significant homogeneity within level 2 units is high. Indeed, in the GCSE English comparability study, it was impossible to fit a model assessing the grading standards at grade F. Convergence was not achieved because, in the population 93.9% of foundation tier candidates were awarded a grade F and, within the sample, 39% of the centres appeared to have no

candidates below the grade F threshold. For other models it was possible to achieve convergence but at the expense of under-dispersion.

A sensible initial step to limit the under-dispersion in comparability study models would be to sample data such that the number of candidates included from each centre was the same. Thereafter efforts would be best focused on increasing the heterogeneity of observed outcomes within centre and ensuring that the probability of achieving such outcomes does not differ between candidates cross-classified identically. Clearly, to be of value, the inter-awarding body comparability studies must focus on the judgmental grade boundaries (See 2.1.1 above) and so there are limits to the scope for increasing the heterogeneity of outcome. It would, however, be possible to restrict analysis to the grade above and below the boundary of interest. This appeals because not only would the observed probability fall somewhere near the centre of the scale but also, by removing candidates at either extreme of the grade distribution, the chances of differing outcome probabilities would be decreased.

Without increasing the sample size, the model for higher tier grade A was recreated selecting only candidates achieving a grade A or B. The estimate of extra-binomial variation was 0.999 with confidence intervals ranging from 0.976 to 1.066. In comparison, for the full model, the confidence intervals ranged from 0.800 to 0.850. The elimination of under-dispersion was at the expense of slight decrease in the predictive efficiency from 0.393 to 0.309 (Pinot de Moira, 2001). A similar pattern was seen across all models: the extent of extra-binomial variation was reduced (although not always completely eliminated) and the predictive efficiency was sometimes compromised.

A natural consequence of restricting the data to include only candidates awarded grades adjacent to the judgmental boundary of interest is a change in the interpretation of the outcome. At the grade A boundary the model would assess whether, as a proportion of those candidates exceeding the grade B boundary, the grade A awards were the same between awarding bodies. To interpret such a model requires an implicit assumption that awarding bodies are competent at placing the grade B boundary; an assumption which is naive given that this boundary is calculated, in part, from the grade A boundary. An inter-awarding body comparability study needs to address the question of whether the proportion of grade A awards over all entries is the same for each awarding body (after controlling for legitimate differences). Indeed, it is possible to create examples where an awarding body may appear to grade leniently when only candidates just above and below the boundary of interest are modelled but severely when all candidates are modelled.

Despite the empirical evidence suggesting that to model only grades adjacent to the judgmental grade boundary under consideration will moderate extra-binomial variation, the interpretative consequences preclude such a methodology. It is therefore recommended that the problems of meeting the binomial assumptions are eased by selecting a dataset such that it is uniform according to Wright's (1997) definition.

2.2.2 Stratified Random Sample or Observational Data, and Issues of Sample Size

Previous inter-awarding body comparability studies have established that the opportunity sample of data available for analysis is often somewhat different in structure from the population. For example, in the GCSE English comparability study the Key Stage 3 match rate differed considerably by grade outcome. Table 1 shows that, to use all the matched data in an analysis would be to over-represent those candidates with grades in the middle of the grade range and to under-represent those at the extremes. Using the same dataset, biases would exist over other subgroups of the population. The Key Stage 3 tests are only compulsory for candidates attending maintained schools and so a model containing all matched data would under-represent independent centres.

TABLE 1 Key Stage 3 match rate by grade

A*	A	B	C	D	E	F	G
31.9%	40.1%	44.2%	48.9%	52.1%	50.3%	43.8%	29.5%

By under-representing certain subgroups of the population, a model will naturally bias towards the subgroups that make up the bulk of the sample as illustrated in Appendix B. Indeed Goldstein (1995) notes 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 …."*

Therefore, to create a model from which between-awarding body grading standards can be validly assessed, the data available for each awarding body must represent the entry for the appropriate syllabus. The volume of data available for use in inter-awarding body comparability studies allows an approach simpler than applying weights. It will generally be possible to sample candidates from the pool of available data such that the resultant dataset is both reasonably representative of each awarding body and large enough to perform a valid analysis. Providing the data represent each awarding body, the number of candidates per awarding body need not represent the market share[3]. Indeed if the data were to represent the market share, then the standard errors associated with the awarding body parameter estimates would be larger for the smaller awarding bodies. The grading standards in the smaller awarding bodies would then be less likely to be singled out as awry. For this reason, while the data should be selected to represent the entry of each individual awarding body, the number of candidates sampled for each awarding body should be the same.

Every year population data are collected from each awarding body to compile the inter-awarding body statistics booklets. These data are cross-classified by syllabus, grade, sex and centre type. Therefore, for each syllabus included in a comparability study, it is possible to determine the population proportions in each of these sub-groups. This information provides a sampling frame from which to draw a stratified random sample of candidates with matched Key Stage 3 results. Obviously this population level classification does not describe exhaustively the differing entry features of the awarding bodies. However, stratifying by grade, sex and centre type, and then selecting candidates randomly but proportionately within these strata, goes some way to redressing the inadequacies of the opportunity sample. In particular, it allows a sufficient proportion of candidates from independent centres to be included in the sample and the proportion of sampled candidates at each grade to reflect the population proportions.

Empirical evidence from the GCSE English comparability study suggests that, when a stratified sample is used to model the probability of exceeding a given grade threshold, the predictive efficiency (Pinot de Moira, 2001) of the model is generally increased. Furthermore, Appendix B suggests that the parameter estimates obtained are less liable to be biased. It seems correct, therefore, that attempts should be made to model data which, as far as possible, reflect the entry of the awarding bodies under consideration. Whether this is best effected by taking a stratified random sample or by re-weighting the matched Key Stage 3 data remains unclear.

[3] Overall population estimates are of no interest in inter-board comparability studies as these data are readily available in inter-awarding body publications.

Certainly by discarding data, as is required when taking the stratified random sample, the precision of the estimates is reduced because the standard errors associated with these estimates increase as the sample size decreases. Modelling the re-weighted opportunity sample would provide parameter estimates with smaller standard errors. In question, therefore, is the appropriate level of precision required to assess the comparability of grading standards between awarding bodies. All other things being equal, the larger the sample, the smaller the difference between awarding bodies which will be deemed statistically significant. Because grade boundaries are placed on a discrete ordinal scale, and marking is completed before determination of grade boundaries, it would be unrealistic to expect that grading standards could be exactly the same between awarding bodies. Indeed, with a large enough sample and small enough mark range, it is possible to conceive of a situation where a one mark increase in the positioning of a grade boundary applied by an awarding body, could mean that the grading standards of that awarding body changed from significantly lenient to significantly severe. The implication of a discrete mark scale to maintenance of grading standards is discussed in more detail by Delap (1992).

As an example, consider a simple model (with no awarding body interactions) fitted to the GCSE English foundation tier grade C candidates to estimate the probability of exceeding the grade C boundary dependent upon awarding body of entry[4]. With an overall sample size of 6,651, the grading of the SEG syllabus appears to be statistically significantly more generous than that of the other award bodies if the grade boundary is placed at 104. As the grade boundary is increased, so the grading of the SEG syllabus becomes less generous, until it becomes statistically significantly more severe if the grade boundary is placed at 109 and higher. According to this simple model therefore, if the grade boundary is placed in the range 105 to 108, in a statistical sense the candidates will not be penalised by the awarding body through which they choose to enter. These 'satisfactory' extremes of grade C boundary would award between 31.3% and 25.8% of candidates a grade C (Table 2). When the sample size decreases, the range of statistically acceptable grade boundaries increases as there is less power to detect a difference (Figure 3). So the optimal sample size must depend upon the level of agreement that can be realistically expected between awarding bodies.

TABLE 2 The proportion of SEG candidates exceeding the grade C boundary dependent upon the chosen boundary mark

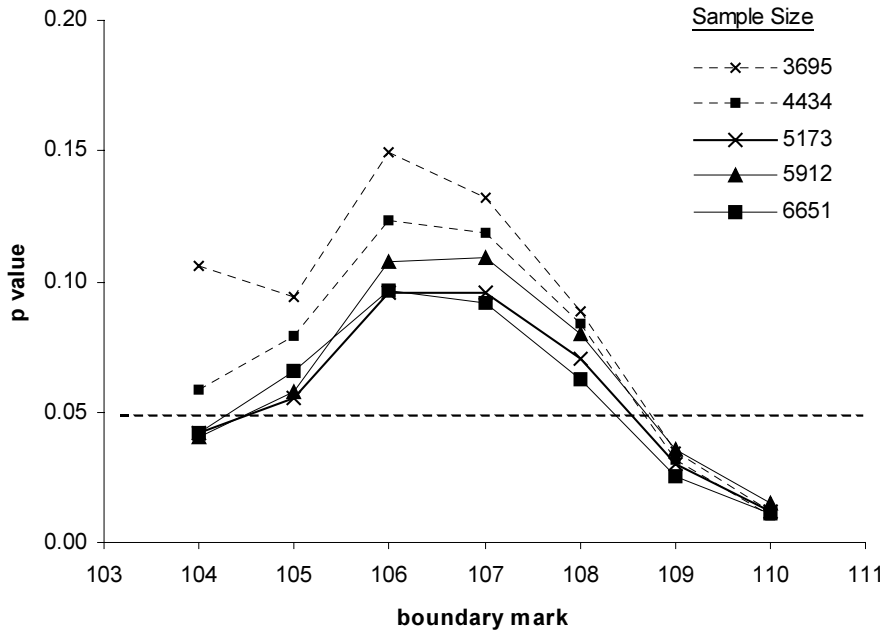
Boundary Mark	104	105	106	107	108	109	110	111
% Grade C (SEG)	32.31	31.28	28.94	27.77	25.79	24.18	22.86	20.73
Difference		1.03	1.34	1.17	1.98	1.61	1.32	2.13

The example in Table 2 shows that a one mark increase in the grade boundary chosen by an awarding committee could result in a change of 2% in the proportion of candidates exceeding that boundary. If the cumulative percentage of grade C candidates for EdExcel fell at, for example, 26.8% then the SEG awarding committee would find it difficult to set the grade C boundary so that the percentage of grade C candidates was less than 1% different from that of EdExcel. Although this limitation is liable to differ between syllabuses dependent upon the maximum mark for the examination and the mean and standard deviation of marks, a 1% difference in the cumulative grade distribution might be described as the most stringent definition of comparable grading standards. At the other extreme, Jones et al (1997) commented that findings from the GCSE Art & Design comparability study were limited by the fact that even a difference in grading standards in excess of half a grade would not be identified as statistically significant. A similar picture emerged from the GCSE English comparability study. In this study, where a main effects model was fitted to the data,

[4] Note that the data used in this example are neither stratified nor re-weighted, they simply serve to illustrate the effect of sample size on the conclusions drawn.

differences of up to 14% in the probability of exceeding the foundation tier grade C boundary were not recognised as statistically significant.

FIGURE 3 **Test of the null hypothesis that there is no difference between the grading standards applied by awarding bodies dependent upon the grade C subject boundary mark applied to the SEG syllabus**



Therefore if, respectively, 1% and 14% are taken as the minimum expectable and maximum acceptable difference between awarding bodies in the percentage of candidates awarded a grade, estimates can be made of the implication these limits have for sample size requirements. The relationship between sample size, significance level, power and effect size and is defined by the following approximation:

$$\frac{\text{Effect Size}\,(\Delta)}{\text{se}(\gamma)} \approx \left(z_{1-\alpha} + z_{1-\beta} \right) \qquad \text{Equation 1}$$

(Snijders & Bosker, 1999)

Where Δ is the difference in log odds between a parameter and the baseline which is detectable from the model.
$\text{se}(\gamma)$ is the standard error of the parameter estimate.

Let $1-\alpha = 0.95 \quad \Rightarrow \quad z_{1-\alpha} = 1.645$
$1-\beta = 0.80 \quad \Rightarrow \quad z_{1-\beta} = 0.842$

Therefore $\text{Effect Size}\,(\Delta) \approx 2.49 \times \text{se}(\gamma)$

To relate the Effect Size (Δ) to the minimum expectable and maximum acceptable difference between awarding bodies, consider the comparison between the parameter estimate for awarding body B compared with the baseline awarding body A.

$$\text{Effect Size}(\Delta) = \gamma_B - \gamma_A = \ln\left(\frac{p+\delta}{1-(p+\delta)}\right) + \ln\left(\frac{p}{1-p}\right) \approx 2.49 \times se(\gamma)$$

Where γ_A is the parameter estimate for awarding body A.

γ_B is the parameter estimate for awarding body B.

p is the probability of exceeding a given grade threshold which is set to 0.5 for the baseline awarding body.

δ is the difference in probability of exceeding a given grade threshold between awarding bodies B and A.

$se(\gamma)$ is now defined as an approximation of the standard error associated with the awarding body parameter estimates.

To create a model where the significance level is 0.05 and the power is 0.80, the $se(\gamma)$ can be estimated in terms of δ and, conversely, δ can be estimated in terms of the $se(\gamma)$.

$$se(\gamma) \approx \frac{\ln\left(\frac{0.5+\delta}{1-0.5-\delta}\right)}{2.49} \qquad \text{Equation 2}$$

$$\delta \approx \frac{e^{2.49\,se(\gamma)} - 1}{2\left(e^{2.49\,se(\gamma)} + 1\right)} \qquad \text{Equation 3}$$

Using an algorithm designed to calculate the power in two-level designs (PINT) and Equation 3 to derive δ, it is possible to estimate the total sample size which will detect a given difference between awarding bodies (Bosker, Snijders, & Guldemond, 1999; Snijders & Bosker, 1993). Total sample size is the product of the number of level 2 units and the number of level 1 units per level 2 unit. For GCSE examinations, there are plenty of candidates entered for each syllabus and it is, therefore, possible to include in excess of sixty candidates per centre. For A Level examinations, each centre may have far fewer entries. The ratio of centres to candidates does have a bearing on the precision with which the parameter estimates are derived. As the number of candidates per centre decreases and the number of sampled centres increases, the nature of the sample tends towards simple random and the precision of the estimates increases (Mok, 1995)[5]. Table 3 presents approximate sample size requirements dependent upon the acceptable difference between awarding bodies and upon the number of candidates per centre. The calculations are based upon several assumptions:

I. α is chosen as 0.05; 1-β is chosen as 0.80.
II. The number of level 1 units per level 2 units is the same for all level 2 units.
III. Each awarding body is equally represented in the dataset.
IV. The model includes the centre level variables: centre type and awarding body and the candidate level variables: mean GCSE achievement and gender.
V. The model includes no interactions.

By defining a simplified model, the standard errors associated with the awarding body parameter estimates may be underestimated and the calculations will yield the minimum sample size necessary to detect a difference of the described magnitude. In other words, for a given sample size, the standard errors reported in Table 3 are liable to be lower than those suggested by the empirical

[5] There is limited additional literature in this area of research. However, Cohen (1998) discusses sample size determination with respect to costs and Afshartous (1995) considers the effect of sample size by referring to empirical evidence.

evidence and, therefore, to maintain the power of the tests, the sample size should be increased from the minimum which is implied by these theoretical calculations.

TABLE 3 Approximate minimum overall sample size requirement needed to detect, as statistically significant, a difference of magnitude δ according to the number of level 1 units per level 2 unit available (α=0.05, 1-β=0.80)

	Overall Sample Size											
	1000		2000		4000		8000		16000		32000	
Level 1 Units	Se(β)	δ	se(β)	δ	se(β)	δ	se(β)	δ	se(β)	δ	se(β)	δ
10	0.236	0.142	0.167	0.102	0.118	0.073	0.083	0.052	0.059	0.037	0.042	0.026
20	0.275	0.164	0.194	0.118	0.137	0.085	0.097	0.060	0.069	0.043	0.049	0.030
30	0.311	0.184	0.220	0.133	0.155	0.095	0.109	0.068	0.077	0.048	0.055	0.034
40	0.340	0.200	**0.240**	**0.145**	0.170	0.104	0.120	0.074	0.085	0.053	0.060	0.037
50	0.368	0.214	0.260	0.156	0.184	0.112	0.130	0.080	0.092	0.057	0.065	0.040
60	0.403	0.231	0.280	0.168	0.198	0.121	0.140	0.086	0.099	0.061	0.070	0.043
70	0.423	0.241	0.299	0.178	0.210	0.128	0.148	0.091	0.105	0.065	0.074	0.046

To return to the example of the maximum acceptable difference between any two awarding bodies (δ=0.14), the theoretical evidence would suggest that to detect such a difference as statistically significant, a sample size of no less than 2,000 candidates would be required with no more than 30 candidates per centre. The dataset from which this maximum acceptable difference was derived comprised 6,651 candidates and, on average, 60 candidates from each centre. The standard errors associated with the awarding body parameter estimates were of the order 0.340. If Table 3 had been used to estimate the minimum difference between awarding bodies which would be deemed statistically significant $\delta \approx 0.09$ might have been assumed. This clearly illustrates that, if any of the assumptions made to calculate the theoretical sample sizes are violated, the power of the tests may be greatly reduced.

A graph of the data in Table 3 is presented in Appendix C & Appendix D provides an annotated example of the output from PINT, detailing how the input parameters were estimated.

In summary, the choice of whether to select a suitably stratified random sample of data or to re-weight the existing data must be based upon whether the ensuing model will be powerful enough to draw meaningful conclusions. In circumstances where the amount of data selected from a stratified random sample would fall short of an sensible minimum, re-weighting should be considered. Whichever way the data are selected, however, the limitations of ignoring the unmatched data should be acknowledged. It is known that these data do not form a random sample of the population. For example, far fewer candidates from independent centres will have a conventional prior measure of achievement for Key Stage 4. Even if, by stratifying or re-weighting, the number of independent centres in the sample is adequate, there is no guarantee that these independent centres represent those for which there is no measure of prior achievement.

2.3 The Independent Variables

2.3.1 Theoretical Considerations in Variable Selection

Matched national data sets are now available and, for many candidates, provide a history of previous qualifications. For A Level candidates, GCSE results are available and for GCSE candidates Key Stage 3 results are available. In fact, since summer 2000 there has been the potential to link longitudinally candidates' progress from Key Stage 3 to A Level. These data have been shown to provide considerable information about candidates' potential. For example, over all the awarding

bodies the correlation between English Key Stage 3 results from summer 1996 and GCSE English results from summer 1998 was 0.74. Furthermore, these matched datasets provide information about concurrent performance. For example, for each candidate sitting GCSE English in 1998, his or her performance in other GCSE examinations is available. With a correlation of 0.86, the concurrent information shows an even greater degree of association with GCSE English grade.

It is certainly an attractive proposition to include such quasi measures of ability into a model assessing the grading standards between awarding bodies. Nevertheless there remain questions over the validity and format of variables that can or should be introduced. The argument underlying the statistical comparability studies is that, after controlling for fair educational achievement, we examine whether there is any unfair or excessive variation in the grades issued by awarding bodies. Although primarily made on the basis of value judgements, selection of the control variables is therefore carried out using statistical criteria as well.

To illustrate this argument, let us select what is probably the least controversial of any of the control variables - centre type. The type of centre that a candidate attends is collected routinely by every awarding body during the processing of an entry. It has long been known that differences are found between the examination performances of candidates from different centre types, so centre type indicators can generally be found that meet the statistical criteria. Let us say that in a specific examination independent centres are awarded better examination grades than other centre types. By including this variable as a control for educational achievement or potential we are implicitly saying that differences between awarding bodies' results can be accounted for by the fact that they have students from different centre types. It could be argued that the question of whether or not it is fair that students from different centre types perform differently in examinations is a matter for society and not one that awarding bodies can control, but awarding bodies have more influence and responsibility than this argument would imply. Decisions about what counts as valuable learning and how that learning is to be assessed are, at the very least, influenced by awarding bodies, although in recent years their influence has been reduced by the government (through the Qualifications and Curriculum Authority). This agenda-setting is, of course, closely tied to the interests of the agenda-setters: for example, witness the introduction of citizenship studies into the curriculum by New Labour. British education tends to be very conservative about change, thereby protecting established societal relationships. Clearly, the awarding bodies cannot control these processes, but neither can they be extricated from them. If we changed to 100% coursework examinations, or excluded Shakespeare from the curriculum, it might change the relationships between the performances of centre types. Would these new relationships be equally fair? Answering this question depends upon our value judgement of what should be assessed and how it should be carried out. This is bound up with our political and cultural values. At least to some extent, the examinations produce the relationships between the examination results for different groups in society, so to use them as controls for differences between the calibre of students entering for different awarding bodies' examinations simply consolidates these relationships. The same argument applies to any of the other variables we may choose to use as controls: gender, ethnic group, free school meals and so on.

If there was a suggestion that examinations were biased against 'poor' children, to what extent would it make sense to say that differences between awarding bodies' standards can be explained by the different numbers of poor children who take their examinations? From another perspective, significant control variables could be interpreted as measures of bias in the examination. The extent to which the relationships measured by the controls are fair depends upon our interpretation, which is closely linked with our value judgements. It is easy to envisage cultural contexts in which our assumptions in the selection of our control variables would be questionable. Future researchers may be outraged that we have included gender as an explanatory variable, arguing that the reason girls did better than boys was due to the feminisation of education (Matters, 1997) and that our comparability studies

allowed this process to go unchecked. Under this view, the most feminised examinations could have been deemed fair by our comparability studies, as their better results were 'justified' by the propensity of female candidates to do better in examinations. Statistical significance does not get us off the hook - examination comparability researchers need to be prepared to defend their selection of fair control variables.

Past discussions, for example, have considered whether or not it is appropriate to include tier of entry as an explanatory variable in the model[6]. The arguments in favour of its inclusion are that the tier into which a candidate is entered provides information about the perceived potential of that candidate. For the purposes of identifying differences in grading standards between awarding bodies, this argument fails if particular awarding bodies are viewed by centres as preferential for particular tiers. There is certainly evidence that this is the case as many centres split their entry between awarding bodies (Baird, 1999; While, 2000b). If there is some reason why centres choose different awarding bodies for different tiers, then to include tier as an explanatory variable in the model is to accept these differences as valid. If centres believe that awarding body A provides an easier assessment of foundation level candidates, then the reasons for choosing the particular tier of entry are no longer simply based upon the perceived potential of the candidate. Under such circumstances, to include tier of entry could be to mask some unacceptable differences in grading standards. This complete argument can be applied equally to the use of estimated grades and there have been similar discussions with respect to the use of common element coursework marks to explain variation in GCSE English grading standards.

The common element coursework comprises 20% of the overall English assessment and, in that it has identical assessment objectives across all awarding bodies, might be assumed to provide a fair measure of candidate ability. Indeed, the correlation between this coursework mark and GCSE English grade was 0.80. However, post hoc analysis suggested that the different moderation techniques, maximum marks and grade boundaries applied by each awarding body, introduced significant variations into the coursework grading (Pinot de Moira, 2000a). As with tier of entry, the inclusion of the coursework mark as an explanatory variable could conspire to mask grading standard differences between awarding bodies.

2.3.2 Statistical Considerations in Variable Selection

The considerations necessary to select explanatory variables do not stop at the point where all controls are determined to be, in some educational sense, fair and valid. In the past, questionnaire data have been collected to augment candidate information that is routinely available. The questionnaires were designed to capture data that were thought to be pertinent to grade outcome in the subjects under consideration (Appendix E). Notwithstanding problems with the data emanating from the questionnaires[7], the introduction to a model of all variables collected from the questionnaires could affect the conclusions drawn from that model. Experience from the GCSE English comparability study showed that the effects measured by the questionnaire variables were swamped by the quasi measure of ability. In other words, the measure of ability was correlated with the questionnaire data, providing evidence of multicollinearity. This phenomenon affects the stability of parameter estimates and increases the associated standard errors such that small changes to the model can lead to large differences in interpretation of the outcome (Kreft & de Leeuw, 1998; Ramsay & Schafer, 1997).

In the context of the comparability studies, the values of individual parameter estimates are of less

[6] Given the discussion and recommendations made in section 2.2.1, it would only be the grade C model where there would be any question over whether to include tier of entry as an explanatory variable.

[7] The overall response rate for the questionnaire was poor, it differed between questions and between different subgroups of the population and there was some doubt over the reliability of the data received.

importance than the comparison between awarding bodies. However, it is possible that the existence of multicollinearity could affect the conclusions drawn regarding grading standards. Certainly if the standard errors associated with the awarding body contrasts were inflated then the probability of wrongly accepting the null hypothesis would also be inflated. Even if there is sound theoretical reasoning behind the collection of questionnaire data, the inclusion into a model of information extracted from the responses could easily be counterproductive if those data explain little more than that which can be gleaned from the quasi measure of ability. Rather than saturating the model with correlated variables to explain as much variation between grade outcomes as possible, variable selection should combine considerations of fairness and validity with those of parsimony.

While it is easy to suggest that questionnaire data have added little to the analysis of grading standards in the comparability studies to date, this does not mean that future analyses should not be augmented by specially collected information. For example, it is regrettable that, given the hierarchical structure under which candidates receive their education, there is currently no routine facility for identifying teacher group. Attempts were made to collect these data from the questionnaires but, because of changes in staffing within centres and a reluctance to divulge information which in the wrong hands could incriminate, there were varying levels of success.

It should also be remembered that, although the quasi measures of achievement explained considerable variation in outcome for the GCSE English examinations, these measures may not be so effective in other subject areas. A comparison between the mean GCSE results in summer 1999 and AS GCE/AS VCE in Summer 2001, shows considerable difference in the correlation (r) and variance explained (r^2) dependent upon certificating subject (Table 4). For example, as might be expected, the correlation between mean GCSE result and AS GCE Chemistry was high whereas that for AS GCE Art & Design was considerably lower. In subjects where an aggregated measure of prior achievement explains a lower proportion of variation in outcome, it may be necessary to collect further information about the characteristics of the entry. Alternatively, it may be possible to formulate the measure of prior achievement more appropriately such that it is not an aggregate of all information. This is discussed further in section 2.3.4.

TABLE 4 The correlation between mean GCSE taken in Summer 2000 and AS GCE/ AS VCE taken in summer 2001

Qualification	r	r^2
AS GCE Chemistry (5421)	0.727	0.529
AS GCE Geography A (5031)	0.719	0.517
AS GCE French (5651)	0.695	0.483
AS GCE Sport & Physical Education (5581)	0.677	0.458
AS GCE English Literature A (5741)	0.662	0.438
AS GCE Mathematics A (5301)	0.648	0.420
AS GCE Drama & Theatre Studies (5241)	0.637	0.406
AS VCE Health & Social Care (8121)	0.623	0.388
AS GCE Business Studies (5131)	0.622	0.387
AS VCE Business (8111)	0.577	0.333
AS GCE Media Studies (5571)	0.542	0.294
AS GCE Art & Design (5201A)	0.513	0.263
AS VCE Information & Communication Technology (8251)	0.484	0.234

2.3.3 To Centre or Not to Centre

Where a raw score is included in a model as an explanatory variable, the interpretation of the intercept derived from that model is the predicted value of the outcome when the raw score is equal to zero. This interpretation might not always be appropriate, particularly if a raw score of zero is outside the range of attainable values. Under such circumstances, it may be more appropriate to centre the raw score around a value which will provide a sensible base from which to interpret the intercept. In a multilevel context, the simplest way to achieve this is to add or subtract the same scalar value from all level 1 units. This is termed grand mean centring and, for models with first order effects only, Kreft & de Leeuw (1998) describe the raw score model and the grand mean centred model as,

> *"…….. equivalent ….. models. This does not mean that all parameter estimates are actually equal. Equivalent models will give the same fit, the same predicted values, and the same residuals, while the parameter estimates can easily be translated into each other."*

Although the terminology suggests that the scalar value for addition or subtraction should be a mean, in practice ease of interpretation is the key motivation for grand mean centring and therefore the scalar value might equally be a median, a mode or some other accepted point from which it would be sensible to drawn conclusions. In the inter-awarding body comparability studies, for example, the Key Stage 3 result was centred around the target level for a fourteen year old candidate.

In models with higher order effects, grand mean centring explanatory variables muddies the comparison between the raw score model and the centred score model. While the parameter estimate for the higher order effect, or interaction, remains the same whether the variables are centred or not, the parameter estimates for the lower order effects differ, not only in value, but in significance. However, Aiken & West (1991) advocate a centred analysis because centred variables provide a more meaningful basis from which to interpret the model outcomes. Subscribing to the convention that when an interaction is included in a model all associated lower order effects should also be included, makes irrelevant the fact that the significance of lower order effects may differ dependent upon whether an explanatory variable is centred or not. The substantive conclusions drawn from the model will not be affected by the centring.

Centring has particular importance in the context of logistic models where the parameter estimates can be compared in terms of odds ratios. By leaving Key Stage 3 level uncentred, for example, the base category would be that of a candidate achieving a level 0 which, in itself, is not a valid outcome. Then ratios for the increased odds of exceeding a given grade threshold would be calculated in relation to the level 0 base measure. It is entirely possible that this would produce sensationally greater odds for candidates at level 8, than for those with the (albeit unachievable) base measure.

Explanatory variables are also sometimes centred within context, or group mean centred[8], such that the level 1 units are centred around the mean of the level 2 unit. Kreft & de Leeuw (1998) recommend that, if the data are group mean centred, the level 2 means should be added as a level 2 variable in addition to the centred level 1 variable otherwise the between level 2 effect will be obscured. Whether to use group mean centring, depends upon the desired way in which the model is to be interpreted. If the aim of inter-awarding body comparability studies is to quantify school[9] effects and to describe candidate effects as deviations from this school effect, then group mean centring (with

[8] As with grand mean centring, for "mean" read mean, mode, median or other appropriate scalar that defines the group around which the level 1 units are centred.

[9] Previously referred to as centres including: schools, colleges and other institutions; but to avoid confusion whilst discussing the centring of variables centres are simply referred to as schools for the remainder of the section.

the school level mean reintroduced) would be appropriate. However, comparability studies are not interested in the schools per se, more in the comparison between awarding bodies. In this context it makes more sense to describe the awarding body parameter estimates in relation to some sensible overall central point.

2.3.4 Variable Format

Even after explanatory variables have been deemed fair in a statistical and theoretical sense, and have been centred where appropriate, there is still some question as to the way in which they should be formulated for inclusion in a model. For example, should candidates' mean GCSE grade or median GCSE grade be used; should the resultant statistic be rounded or truncated; should only a subset of GCSEs be used to create this statistic? Where centre type is to be included, should each of the eleven centre types be modelled or should the variable be collapsed to distinguish between, for example, selective and non-selective centres? Where an ordinal variable is to be introduced should it be treated as discrete by creating a set of binary contrasts or would it suffice to treat it as continuous?

With the exception of the awarding body contrasts, the independent variables are included for their capacity to explain variation, not to provide specific information about their affect upon grade outcome. So, unless the way in which these variables are formulated alters the extent to which they explain the variation, then it could be argued that such considerations are irrelevant. However, in a paper written for the AQA Alliance Standards Unit, While (2000a) alluded to the fact that variable formulation can affect significance. In an extension to this analysis, Table 5 shows a differing correlation between GCSE grade and various measures of concurrent achievement. What is interesting is that no matter what the subject, the unadulterated mean GCSE grade has the highest correlation with the subject-specific GCSE grade outcome (see the emboldened cells). Furthermore, when measures of prior achievement are considered, in the form of Key Stage 3 results, it is the unadulterated mean of the three tests that has the highest correlation with GCSE outcome for most subjects (see the emboldened cells in Table 6). With GCSE Mathematics the best predictor appears simply to be the Key Stage 3 level achieved in the same subject area. Nevertheless this statistic is only marginally better than the mean statistic.

TABLE 5 Correlation between various independent variables and GCSE grade outcome for a selection of GCSE subjects taken in Summer 1998

	Geography	English	Double Science	History	Maths	Business Studies	French	PE & Sport	Art & Design
Mean GCSE	**0.926**	**0.924**	**0.917**	**0.914**	**0.904**	**0.892**	**0.878**	**0.804**	**0.796**
Truncated Mean GCSE	0.911	0.911	0.902	0.899	0.890	0.879	0.862	0.787	0.784
Rounded Mean GCSE	0.911	0.910	0.902	0.898	0.891	0.872	0.863	0.786	0.787
Median GCSE	0.910	0.904	0.911	0.895	0.887	0.862	0.851	0.771	0.762
Mean English, Maths & Science	0.891	-	-	0.868	-	0.833	0.827	0.727	0.721
% GCSE grades >=C	0.884	0.888	0.880	0.857	0.866	0.862	0.842	0.781	0.782
Categorised Mean GCSE[10]	-0.882	-0.862	-0.843	-0.885	-0.837	-0.809	-0.836	-0.691	-0.723
Sum of GCSE grades	0.879	0.892	0.882	0.859	0.859	0.864	0.829	0.782	0.762
% GCSE grades >=A	0.790	0.772	0.714	0.817	0.725	0.670	0.751	0.633	0.693
Selective/Non-Selective	0.466	0.468	0.340	0.493	0.474	0.128	0.493	0.187	0.375
Selective/Non-Selective/FE	0.465	0.464	0.339	0.492	0.472	0.128	0.493	0.184	0.373
Centre Type	0.319	0.340	0.241	0.404	0.343	0.031	0.368	0.083	0.379
Number of entries	0.232	0.380	0.331	0.175	0.380	0.306	0.235	0.337	0.361

[10] Candidates were divided among ten groups using the same categories as determined to predict summer 2001 AS results. If the mean GCSE was greater than 7.10 then they were in group 1, greater than 6.69 group 2, greater than 6.38 group 3, greater than 6.11 group 4, greater than 5.89 group 5, greater than 5.62 group 6, greater than 5.36 group 7, greater than 5.09 group 8, greater than 4.69 group 9 and otherwise group 10.

TABLE 6 Correlation between Key Stage 3 variables and GCSE grade outcome for a selection of GCSE subjects taken in Summer 1998

	Geography	English	Double Science	History	Maths	Business Studies	French	PE & Sport	Art & Design
Mean Key Stage 3	**0.801**	**0.815**	**0.807**	**0.765**	0.842	**0.718**	**0.743**	**0.644**	**0.615**
Truncated Mean Key Stage 3	0.770	0.777	0.768	0.733	0.807	0.700	0.707	0.599	0.564
Rounded Mean Key Stage 3	0.766	0.776	0.768	0.729	0.807	0.688	0.705	0.598	0.564
Sum of Key Stage 3	0.705	0.735	0.718	0.689	0.754	0.648	0.651	0.580	0.561
Mathematics Key Stage 3	0.734	0.712	0.763	0.682	**0.858**	0.656	0.664	0.612	0.545
English Key Stage 3	0.654	0.758	0.583	0.658	0.611	0.566	0.645	0.480	0.563
Science Key Stage 3	0.722	0.692	0.773	0.679	0.753	0.630	0.646	0.580	0.513

Tentatively, we may conclude that with measures of prior and concurrent achievement the best formulation for an explanatory variable might be an unrounded and untruncated mean. Certainly this formulation seems to be better, in all circumstances explored, than other variables created by combining information from all qualifications. It may be that, for some subjects, a subset of previously or concurrently gained qualifications may provide a stronger relationship with a particular GCSE outcome. The decision to explore subsets of the concurrent or prior achievement data would, however, need to be informed by theoretical reasoning (see section 2.3.1) to avoid mindless data dredging.

Table 5 also demonstrates the relationship between centre type and GCSE grade outcome in a particular subject. What is interesting is the extent to which the formulation of this variable affects correlation. When disaggregated to distinguish between eleven centre types, it is a less powerful predictor of grade outcome than when summarised. In the subjects considered in Table 5 there is nothing to be gained by making any distinction between centre types except to qualify whether they are selective or non-selective. It should be remembered, however, that GCSE subjects are entered predominantly by candidates in schools. With future inter-awarding body comparability studies which may investigate post sixteen qualifications, a binary classification of centre may be of less value than some alternative and theoretically appropriate formulation.

Measures of correlation such as those reported in Table 5 & Table 6 provide a crude statistical measure by which to assess the most appropriate formulation of variables. They suggest that the aggregate of prior or concurrent achievement is best left in its least refined form with no rounding truncating or categorising. Such a finding naturally leads to the question of whether it is more appropriate to include ordinal discrete data into a model as a set of binary contrasts or as a continuous variable. Clearly there are no hard and fast rules which would be equally applicable to all situations. The decision rests upon, among other things, the fineness of the ordinal scale. However, evidence from the GCSE Mathematics inter-awarding body comparability study (While & Fowles, 2000) suggests that, where ordinal variables were included as categorical, the parameter estimates associated with the increasing values could, with some imagination, be described as monotonically linearly increasing. For this reason it could be concluded that, where inferences were not required of the parameter estimates, it would be sufficient to treat ordered discrete variables as continuous.

3 CONCLUSIONS & RECOMMENDATIONS

Current practice in the statistical assessment of inter-awarding body comparability is to fit a logistic model to determine the probability of exceeding a given grade threshold. An ordered multinomial model would also provide a satisfactory tool to model grading standards and to suggest necessary remedial action. However, the complex nature of the ordered multinomial model means that currently it is more computationally intensive to fit and more difficult to interpret. In future inter-awarding body comparability studies it is, therefore, recommended that:

- A logistic model should continue to be fitted to assess the probability of exceeding a given grade boundary.

A considerable number of the issues discussed with respect to the statistical modelling of grading standards, are focussed on the refinement of standard errors associated with the model parameter estimates, particularly those for the awarding bodies. In terms of statistical significance, the accurate evaluation of standard errors is important to minimise the probability of Type I and Type II errors. While the implementation of some recommendations would reduce the standard errors and therefore reduce the possibility of erroneously concluding there is no difference between awarding bodies, some would actually increase the standard errors. For example, after considering the arguments for and against the modelling of an inherent hierarchy, it was concluded that to treat individual candidates as completely independent when they are nested within centres would be wrong. The effect of modelling the hierarchy is to increase the standard errors associated with the parameter estimates by conceding that the effective sample size is smaller than that suggested by the number of level one units. Within the multilevel framework, other measures recommended to increase the robustness of the analysis would reduce the standard errors associated with the parameter estimates. For example, the recommendation to minimise multicollinearity between independent variables would serve to decrease standard errors and reduce the probability of erroneously concluding there is no difference between awarding bodies.

Although the cumulative effects of these recommendations appear to cancel each other out by both increasing and decreasing the standard errors simultaneously, together they improve the accuracy with which the standard errors are estimated. In future inter-awarding body comparability studies, therefore, the following measures should be put in place:

- The inherent hierarchy should be reflected in a multilevel model.
- The number of level 1 units per level 2 unit should be the same for all level 2 units.
- Independent variables must be chosen carefully to avoid the adverse effects of multicollinearity.

The standard errors associated with parameter estimates are also affected by the sample size and sample design. The larger the sample the smaller the standard errors. Furthermore, predictive efficiency appears to increase when the sample more adequately represents the entry for each awarding body. The recommendations suggest that the computationally simpler strategy of selecting a stratified random sample from the pool of matched data should be employed only when enough data are available such that the ensuing model will be powerful enough to draw meaningful conclusions. Otherwise the application of weights to the dataset which reflect awarding body entries should be considered. In future inter-awarding body comparability studies, therefore, the data should be selected as follows:

- A stratified random sample of the data should be used to model grading standards. At the very least, the strata should include grade, sex and centre type[11].
- For each judgmental boundary, the sample of data should include candidates across the full grade range rather than just candidates achieving grades adjacent to the judgmental boundary of interest.
- Minimum sample size requirements should be estimated from the graphs included in Appendix C.

The expounding of issues surrounding the choice and formulation of independent variables does not produce such clear cut recommendations. In the context of inter-awarding body comparability studies

[11] Attempts to achieve this recommendation may preclude successfully selecting a uniform number of level 1 units per level 2 unit. However, selecting a stratified random sample should be regarded as a higher priority.

it seems sensible to centre variables around some grand mean so that the awarding body parameter estimates can be interpreted with reference to some meaningful baseline. Independent variables should be introduced to the model based upon both value judgements and with a view to the statistical implications such as the issues of multicollinearity. Independent variables will not be equally informative across all subject areas for which comparison of standards might be needed. In some circumstances the readily available data may need to be reconfigured to suit the purpose or alternatively it may be necessary to augment these data with information collected from other sources.

4 ACKNOWLEDGMENTS

With grateful thanks to Jo-Anne Baird for her help with the arguments presented in section 2.3.1 and to Lesley Meyer for her help with section 2.2.1 & 2.2.2.

5 REFERENCES

Afshartous, D. (1995). *Determination of sample size for multilevel model design*. Paper presented at the AERA meeting, San Francisco.

Aiken, L., & West, S. (1991). *Multiple regression: Testing and interpreting interactions*. London: Sage Publications.

Baird, J. (1999). Standards in GCSE Modern Foreign Languages: statistical comparisons of the SEG and NEAB examinations. *Internal Alliance Standards Unit Report*.

Baird, J., Cresswell, M., & Newton, P. (1999). Would the real gold standard please step forward? *RC Paper*, *RC/22*.

Baird, J., & Jones, B. E. (1998). Statistical Analyses of Examination Standards: Better measures of the unquantifiable? *RAC Paper*, *RAC/780*.

Barcikowski, R. S. (1981). Statistical power with group mean as the unit of analysis. *Journal of Educational Statistics*, *6*(3), 267-85.

Bell, J. (1999). *Initial analysis of English Literature data*. Paper presented at the Methodologies of Recent Comparability Studies Seminar, Manchester.

Bosker, R. J., Snijders, T. A. B., & Guldemond, H. (1999). PINT (Power IN Two-level designs): Estimating standard errors of regression coefficients in hierarchical linear models for power calculations, User's Manual. .

Cohen, M. P. (1998). Determining the sample sizes for surveys with data analyses by hierarchical linear models. *Journal of Official Statistics*, *14*(3), 267-275.

Delap, M. R. (1992). Statistical Information At Awarding Meetings: The Discrete Nature Of Mark Distributions. *RAC Paper*, *RAC/585*.

Fielding, A. (1999). Why use arbitrary points scores?: ordered categories in models of educational progress. *Journal of the Royal Statistical Society Series A (Statistics in Society)*, *162*(3), 303-328.

Fitz-Gibbon, C. T. (1997). *The value added national project final report: Feasibility studies for a national system of value-added indicators* . The School Curriculum and Assessment Authority (SCAA).

Fowles, D. (2000). A review of the methodologies of recent comparability studies: Report on an inter-board staff seminar. *RC Paper*, *RC/42*.

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

Hox, J. (1995). *Applied Multilevel Analysis*. Amsterdam: TT-Publikaties.

Jones, B., Baird, J., & Arlett, S. (1997). A comparability study in GCSE Art & Design (Unendorsed): A study based on the Summer 1996 examinations. *Organised by the NEAB on behalf of the Joint Forum for the GCSE and GCE*.

Kreft, I. (1996). Are multilevel techniques necessary? An overview including simulation studies. *www.stat.ucla.edu/~kreft/quarterly/quarterly.html*.

Kreft, I., & de Leeuw, J. (1998). *Introducing multilevel modelling*. (First ed.). London: SAGE Publications Ltd.

Matters, G. (1997). *Are Australian boys underachieving? An analysis using the Cronbach-Moss Validity-Reliability framework*. Paper presented at the International Association for Educational Assessment conference, Durban, South Africa.

Mok, M. (1995). Sample size requirements for 2-level designs in educational research. *Multilevel Modelling Newsletter*, *7*(2), 11-15.

Pinot de Moira, A. (2000). *A Comparability Study in GCSE English: Statistical analysis of results by board* . A study based on the Summer 1998 examination and organised by AQA(SEG) on behalf of the Joint Forum for the GCSE and GCE.

Pinot de Moira, A. (2001). Explained variance in logistic multilevel models. *RC Paper*, *RC/119*.

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

Snijders, T. A. B., & Bosker, R. J. (1993). Standard errors and sample sizes in two level research. *Journal of Educational Statistics*, *18*(3), 237-260.

While, D. (2000a). Brief (A) preliminary report to summarise the work-in-progress of the A level Delta Analyses using the Matched 16+/18+ national database. *RC Paper*, *RC/60*.

While, D. (2000b). Subjects with centres operating a 'split entry policy' - a report produced for the ASU. *RC Paper*, *RC/69*.

While, D., & Fowles, D. (2000). *A Comparability Study in GCSE Mathematics - Statistical analysis of results by board* . A Study based on the Summer 1998 examination and organised by AQA(NEAB) on behalf of the Joint Forum for the GCSE and GCE.

Wright, D. B. (1997). Extra-binomial variation in multilevel logistic models with sparse structures. *British Journal of Statistical and Mathematical Psychology*, *50*(1), 21-30.

APPENDIX A DEPENDENT VARIABLE CHOICE

A linear, logistic and ordered multinomial model have been fitted to the same higher tier GCSE English data from summer 1999 (Model 1 - Model 3, described overleaf). The independent variables included in each model are essentially the same and were selected simply to illustrate the implications of choosing different dependent variables. Each model of grade outcome includes, as a covariate, the level achieved in Key Stage 3 English and, as a factor, a contrast between the grading of the NEAB legacy syllabus and that of syllabuses offered by other awarding bodies. The conclusions that may be drawn from each of models with respect to the grading of NEAB candidates are summarised in Table 7.

TABLE 7 Interpretation of the parameter estimates (predictions for a candidate achieving English Key Stage 3 Level 5)

Model	Interpretation
1 Linear	A candidate entering the NEAB examination is predicted to gain 0.143 of a grade less than his/her counterpart entered through another awarding body. In other words, the award of the NEAB syllabus is, as an average across the whole grade range, $^1/_7$ of a grade more severe than that of other awarding bodies.
2 Logistic	The cumulative proportion of NEAB candidates predicted to be awarded a grade C is $\frac{e^{(2.476-0.794)}}{1+e^{(2.476-0.794)}} * 100 = 84.3\%$. The cumulative proportion of candidates entered through other awarding bodies predicted to be awarded a grade C is $\frac{e^{(2.476)}}{1+e^{(2.476)}} * 100 = 92.2\%$.
3 Ordered Multinomial	The cumulative proportion of NEAB candidates predicted to be awarded a grade C is $\frac{e^{(1.917-0.274-0.303)}}{1+e^{(1.917-0.274-0.303)}} * 100 = 79.2\%$. The cumulative proportion of candidates entered through other awarding bodies predicted to be awarded a grade C is $\frac{e^{(1.917)}}{1+e^{(1.917)}} * 100 = 87.2\%$.

Differences between the predictions are a function of the model fitted. Even with the logistic and ordered multinomial models, where the predictions are expressed in the same manner, the cumulative percentages are affected differentially by the decision either to model a binary contrast or to model each grade individually. From the interpretation presented in Table 7 the latter two models suggest that, for candidates with an average English Key Stage 3 result, the NEAB syllabus allows approximately 8% fewer candidates to achieve a grade C than the syllabuses offered by other awarding bodies[12]. This information could easily be used in award meetings given the nature of the data available at these meetings. In contrast, the information presented by the linear model does not provide the tools to effect remedial action. The model does not locate the problem at a particular grade boundary, nor does it express the magnitude of differences in terms of the cumulative percentage measurement used to compare awards.

[12] The findings reported in these examples exclude many significant independent variables and should not therefore be taken to indicate grading inconsistencies between any of the syllabuses under consideration.

MODEL 1 Linear Model

$$\text{gcseeng}_{ij} \sim \text{N}(XB,\ \Omega)$$

$$\text{gcseeng}_{ij} = \beta_{0j}\text{cons} + 0.642(0.011)\text{eng}_{ij} + -0.143(0.070)\text{neab}_j$$

$$\beta_{0j} = 0.280(0.037) + u_{0j} + e_{0ij}$$

$$\begin{bmatrix} u_{0j} \end{bmatrix} \sim \text{N}(0,\ \Omega_u) \ :\ \Omega_u = \begin{bmatrix} 0.102(0.015) \end{bmatrix}$$

$$\begin{bmatrix} e_{0ij} \end{bmatrix} \sim \text{N}(0,\ \Omega_e) \ :\ \Omega_e = \begin{bmatrix} 0.613(0.010) \end{bmatrix}$$

MODEL 2 Logistic Model

$$\left.\begin{aligned} &\text{pgradc}_{ij} \sim \text{Binomial}(\text{denom}_{ij},\ \pi_{ij}) \\ &\text{pgradc}_{ij} = \pi_{ij} + e_{0ij}\text{bcons}^{*} \end{aligned}\right\}$$

$$\text{logit}(\pi_{ij}) = \beta_{1j}\text{cons} + 1.473(0.063)\text{eng}_{ij} + -0.794(0.265)\text{neab}_j$$

$$\beta_{1j} = 2.476(0.145) + u_{1j}$$

$$\begin{bmatrix} u_{1j} \end{bmatrix} \sim \text{N}(0,\ \Omega_u) \ :\ \Omega_u = \begin{bmatrix} 1.227(0.215) \end{bmatrix}$$

$$\text{bcons}^{*} = \text{bcons}[\pi_{ij}(1 - \pi_{ij})/\text{denom}_{ij}]^{0.5}$$

$$\begin{bmatrix} e_{0ij} \end{bmatrix} \sim (0,\ \Omega_e) \ :\ \Omega_e = \begin{bmatrix} 1.000(0.000) \end{bmatrix}$$

MODEL 3 Ordered Multinomial Model

$$\text{cumu}_{ijk} \sim \text{N}(XB,\ \Omega)$$

$$\begin{aligned} \text{cumu}_{ijk} = {} & -5.072(0.113)\text{a}^{*}_{ijk} + -2.707(0.093)\text{a}_{ijk} + -0.607(0.088)\text{b}_{ijk} + \\ & 1.917(0.093)\text{c}_{ijk} + 5.093(0.166)\text{d}_{ijk} + 1.481(0.030)\text{eng}_{ijk} + \\ & -0.274(0.165)\text{neab}_k + -0.015(0.079)\text{a\_neab}_{ijk} + \\ & -0.303(0.088)\text{c\_neab}_{ijk} + v_{7k}\text{cons} \end{aligned}$$

$$\begin{bmatrix} v_{7k} \end{bmatrix} \sim \text{N}(0,\ \Omega_v) \ :\ \Omega_v = \begin{bmatrix} 0.534(0.080) \end{bmatrix}$$

APPENDIX B SAMPLE BIASES

Consider the case where a model of the probability of achieving a grade C is produced which includes awarding body and candidate sex as explanatory variables. For awarding bodies X & Y, there are plenty of data relating to both male and female candidates. For awarding body Z there are few data relating to females. Both the main effects and the awarding body-sex interaction are fitted to the data but the neither of the components of the interaction are statistically significant because the relationship between grade outcome and sex appears to be the same across all awarding bodies (Model 4).

MODEL 4 Biased sample

Distribution of male and female candidates between awarding bodies

	X	Y	Z	Total
Male	1470	1616	1882	4968
Female	1338	1535	58	2931
Total	2808	3151	1940	7899

$$\left.\begin{aligned} \text{pgradc}_{ij} &\sim \text{Binomial}(\text{denom}_{ij}, \pi_{ij}) \\ \text{pgradc}_{ij} &= \pi_{ij} + e_{0ij}\text{bcons}^{*} \end{aligned}\right\}$$

$$\begin{aligned} \text{logit}(\pi_{ij}) = {} & 0.267(0.047)\text{cons} + 0.299(0.277)\text{female}_{ij} + \text{-}0.046(0.070)\text{x}_{j} + \\ & \text{-}0.146(0.068)\text{y}_{j} + 0.335(0.288)\text{female.x}_{ij} + 0.431(0.287)\text{female.y}_{ij} \end{aligned}$$

$$\text{bcons}^{*} = \text{bcons}[\pi_{ij}(1 - \pi_{ij})/\text{denom}_{ij}]^{0.5}$$

$$\left[e_{0ij}\right] \sim (0, \Omega_{e}) : \Omega_{e} = \left[1.000(0.000)\right]$$

However, the estimates of the awarding body-sex interaction are dominated by the relationship seen in awarding bodies X & Y. By collecting more information about the female candidates entered through awarding body Z, the refitted model shows that there is in fact a considerable difference in the male/female performance between awarding bodies (Model 5). Male candidates from awarding body Z have been set as the baseline. The negative parameter estimate associated with female performance suggests that the probability of female candidates from awarding body Z exceeding the grade C threshold is lower than that for male counterparts. The positive awarding body-sex interaction parameter estimates show that the pattern of achievement for female candidates entered through awarding bodies X & Y is quite the opposite. In these two boards female candidates do much better.

Clearly, by under-representing certain subgroups of the population in an analysis, there is an increased chance that the conclusions drawn about that subgroup will be incorrect. In addition, population estimates derived from the model will be biased in favour of the represented groups.

MODEL 5 Stratified sample

Distribution of male and female candidates between awarding bodies

	A	B	C	Total
Male	1470	1616	1940	5026
Female	1338	1535	2054	4927
Total	2808	3151	3994	9953

$$\left.\begin{aligned} \text{pgradc}_{ij} &\sim \text{Binomial}(\text{denom}_{ij}, \pi_{ij}) \\ \text{pgradc}_{ij} &= \pi_{ij} + e_{0ij}\text{bcons}^{*} \end{aligned}\right\}$$

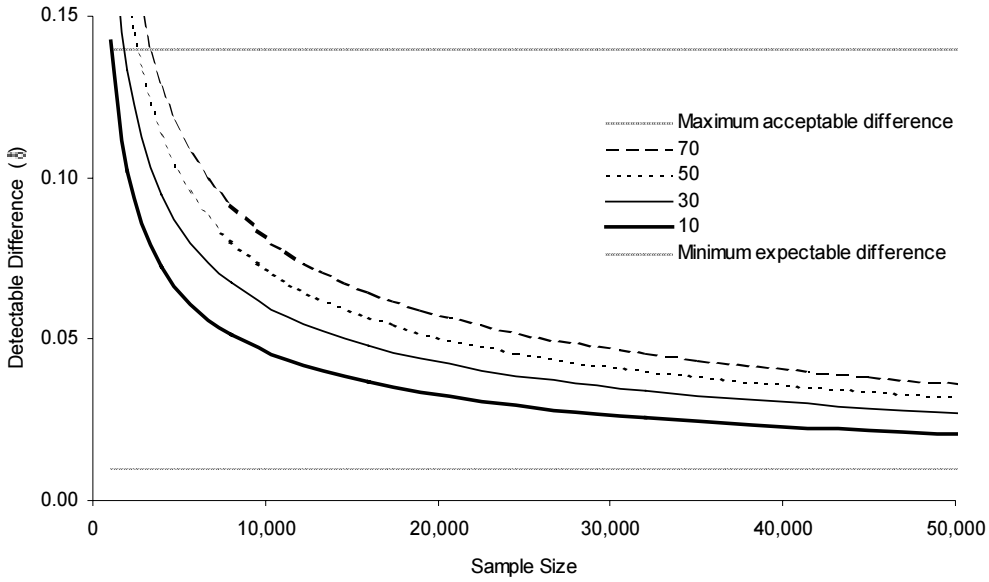
$$\begin{aligned} \text{logit}(\pi_{ij}) = {} & 0.276(0.046)\text{cons} + \text{-}0.666(0.064)\text{female}_{ij} + \text{-}0.055(0.070)\text{x}_{j} + \\ & \text{-}0.155(0.068)\text{y}_{j} + 1.301(0.102)\text{female.x}_{ij} + 1.397(0.099)\text{female.y}_{ij} \end{aligned}$$

$$\text{bcons}^{*} = \text{bcons}[\pi_{ij}(1 - \pi_{ij})/\text{denom}_{ij}]^{0.5}$$

$$\left[e_{0ij}\right] \sim (0, \Omega_e) : \Omega_e = \left[1.000(0.000)\right]$$

APPENDIX C SAMPLE SIZE REQUIREMENTS

FIGURE 4 Minimum overall sample size requirement needed to detect, as statistically significant, a difference of the magnitude δ dependent upon number of level 1 units per level 2 unit (α=0.05, 1-β=0.80)



APPENDIX D ANNOTATED PINT OUTPUT

```
  PPPPPP       III    NNN      NNN  TTTTTTTTTTTTT
  PPP   PPP         NNNN     NNN       TTT
  PPP   PP     III    NNN NN   NNN       TTT
  PPPPPP       III    NNN  NN  NNN       TTT
  PPP          III    NNN   NN NNN       TTT
  PPP          III    NNN    NNNNN       TTT
  PPP          III    NNN     NNNN       TTT
  PPP          III    NNN      NNN       TTT

           power in two-level designs
           version 1.61       april 1999
copyright:   roel bosker    (UT)  & tom snijders (RUG)
programming: henk guldemond (RUG) & tom snijders (RUG)

This programs performs calculations corresponding to the paper
"Standard errors and Sample Sizes for Two-Level Research",
by Tom A.B. Snijders and Roel J. Bosker,
Journal of Educational Statistics, Vol. 18, 1993, p. 237-259.

date, d-m-y 26 -  9 - 2001
hour, h:m:s 14 : 13 :   56.53
Input read from file ax.dat.

Design:
(Between parentheses, the symbol is mentioned that is used for this parameter;
 if the PINT manual uses a different symbol than the Snijders-Bosker paper,
 then the symbol from the paper is given between parentheses,
 the symbol from the manual between square brackets.)

NUMBER OF FIXED EFFECTS             (K_1)            [L_1 + L_2 + 1] :    2   ⇒ Mean GCSE & Gender
NUMBER OF RANDOM EFFECTS INCL. CONST (K_2)           [L_2 + 1]       :    1   ⇒ Constant
NUMBER OF LEVEL-2 VARS INCL. CONST   (length of W_3j) [L_3 + 1]      :    6   ⇒ Constant + centre type +
TOTAL COSTS                          (K)                             : 2000       4 awarding body contrasts
RELATIVE COST PER LEVEL-2 UNIT       (n)                             :    0
SMALLEST VALUE OF n                                   [n_min]        :   10
STEP SIZE FOR n                                       [n_step]       :   10
LARGEST VALUE FOR n                                   [n_max]        :   70

Parameters:

WITHIN-GROUPS COVARIANCE MATRIX  (SIGMA-W)

      0.25000     -0.01000      0.00000
     -0.01000      0.90000      0.00000
      0.00000      0.00000      0.00000

BETWEEN-GROUPS COVARIANCE MATRIX (SIGMA-B)

0.00000      0.00000      0.00000      0.00000      0.00000      0.00000      0.00000      0.00000
0.00000      0.03000     -0.00100     -0.00100     -0.00100     -0.00100     -0.00100     -0.00100
0.00000     -0.00100      0.16000     -0.04000     -0.04000     -0.04000     -0.01000     -0.01000
0.00000     -0.00100     -0.04000      0.16000     -0.04000     -0.04000     -0.01000     -0.01000
0.00000     -0.00100     -0.04000     -0.04000      0.16000     -0.04000     -0.01000     -0.01000
0.00000     -0.00100     -0.04000     -0.04000     -0.04000      0.16000     -0.04000     -0.01000
0.00000     -0.00100     -0.01000     -0.01000     -0.01000     -0.04000      0.03000     -0.01000
0.00000     -0.00100     -0.01000     -0.01000     -0.01000     -0.01000     -0.01000      0.06000
```

⇒ *Covariance matrix for level 1 variables. Ignore zeros in row & column 3. Top left: estimate of the variance in gender at level 1 alone. Calculated as the variance in the difference, for each candidate, from the "mean gender" for their centre.. Bottom right: similarly calculated candidate level variance in mean GCSE. Off diagonals: covariance between the two.*

⇒ *Covariance matrix for level 2 variables. Ignore zeros in row & column 1. Variables ordered as follows: Centre type, 4 awarding body contrasts (note Variance and covariances are the same for all of these), centre level gender effect, centre level mean GCSE. The leading diagonal contains the variances and the off diagonals the covariances.*

Estimates of the values in these covariance matrices were made from empirical examples. For the level 1 variables the key is to split the variation between level 1 and level 2. As mentioned above, this is best effected by group mean centring the variables to find the level 1 variation and, at level 2, by calculating the variance between group means.

```
RESIDUAL VARIANCE (sigma-squared)

          3.29000
```

⇒ *This is a logistic model and therefore the residual variance at level 1 is constrained to equal* $\pi^2/3$.

COVARIANCE MATRIX OF RANDOM EFFECTS (tau_2) ⇒ ***Residual variance at level 2 is estimated from empirical evidence provided by previous models fitted.***

0.20000

EXPECTATION OF LEVEL-1 VARIABLES WITH FIXED EFFECTS (mu_1)

0.50000 ⇒ ***Mean gender at level 1.***
-0.80000 ⇒ ***Mean GCSE result at level 1***

EXPECTATION OF LEVEL-2 VARIABLES (mu_3)

1.00000 ⇒ ***Mean value of the constant.***
0.03000 ⇒ ***Mean value of centre type.***
0.20000 ⇒ ***Mean value of awarding body 1.***
0.20000 ⇒ ***Mean value of awarding body 2.***
0.20000 ⇒ ***Mean value of awarding body 3.***
0.20000 ⇒ ***Mean value of awarding body 4.***

CONSTANT MEAN VECTOR (mu_2 = e)

1.00000

The following table contains the standard errors (s.e.):
Fixed: s.e. of regr. coeff.s of level-1 variables with a fixed effect only.
Const: s.e. of the intercept.
Group: s.e. of regr. coeff.s of level-2 variables.
Random: s.e. of regr. coeff.s of level-1 variables with a random effect.
Cross-L: s.e. of regr. coeff.s of cross-level interactions
(product of "Group" with "Random effect" variables).

Sample sizes			Standard errors							
N*n	N	n	Fixed	Fixed	Const	Group	Group	Group	Group	Group
2000	200	10	0.08168	0.04210	0.13693	0.29754	0.16654	0.16654	0.16654	0.16943
2000	100	20	0.08152	0.04227	0.15445	0.34925	0.19429	0.19429	0.19429	0.19675
1980	66	30	0.08184	0.04259	0.17100	0.39622	0.21965	0.21965	0.21965	0.22185
2000	50	40	0.08138	0.04244	0.18450	0.43458	0.24038	0.24038	0.24038	0.24236
2000	40	50	0.08134	0.04249	0.19781	0.47149	0.26038	0.26038	0.26038	0.26221
1980	33	60	0.08172	0.04274	0.21134	0.50827	0.28036	0.28036	0.28036	0.28208
1960	28	70	0.08212	0.04298	0.22430	0.54323	0.29938	0.29938	0.29938	0.30101
			Gender	***Mean Gcse***	***Constant***	***Centre Type***	***AB1***	***AB2***	***AB3***	***AB4***

APPENDIX E QUESTIONNAIRE

GCSE English (*Subject Code*)
Candidate Questionnaire

The GCSE Examining Groups are carrying out a study to make sure that the GCSE English examinations in each Group are the same standard. To make a fair comparison, we need to know a little about you and your fellow students. The following questionnaire asks about you and your parents, and about how you feel about GCSE English.

All of the information given will be treated with strict confidentiality and anonymity.

1. Please print your name in capital letters on the line below.

.....

2. Are you male or female? ❑ male
Please tick a box ❑ female

3. What is your date of birth?
Please enter the day, month and year day : month : year

4. Please print in capital letters the name of the teachers (or teacher) who taught your English class this year.

.....
.....
.....

5. How much do you like school/college?
Please circle the number that best describes how you usually feel about school/college.

1 I really enjoy school/college.
2 I like school/college.
3 I neither like nor dislike school/college.
4 I do not like school/college.
5 I really dislike school/college.

6. How much do you like English as a subject?
Please circle the number that best describes how you usually feel about English.

1 I really enjoy English.
2 I like English.
3 I neither like nor dislike English.
4 I dislike English.
5 I really dislike English.

7. How many hours do you spend on English homework each week?

Please write the number of hours: .....hours
(write '0' if you do not usually do any homework)

8. Do you intend to sit the following kinds of examinations in the next two years?
Please tick as many boxes as you need to.

- ❑ GCSEs
- ❑ NVQs
- ❑ A levels
- ❑ GNVQs
- ❑ other: *please state which* .....
- ❑ none

9. What do you intend to do when you leave school/college?
Please tick the box next to the statement that best describes what you think you will do when you leave school/college.

- ❑ Try to find a job.
- ❑ Go to University and do a course in English.
- ❑ Go to University and do a course in a subject other than English.
- ❑ None of the above

10. Are you entitled to receive free school meals? ❑ Yes
Please tick a box ❑ No

11. How encouraged are you at home to study English?
Please circle the number next to the statement that best describes you

1 I am very much encouraged to study English.
2 I am encouraged a little to study English.
3 I am neither encouraged nor discouraged to study English.
4 I am discouraged a little from studying English.
5 I am very much discouraged from studying English.

12. What was your Key Stage 3 result for English?
Please write the level you got on the line. If you cannot remember the level, your teacher might be able to help. Otherwise, just leave the line blank and go on to the next question.

level _____

13. What is your home post code? *Please write it in the boxes*

Thank you for completing the questionnaire. Please return it to your teacher.