**Paper for Fifth International Conference on Social Science Methodology of the Research Committee on Logic and Methodology (RC33) of the International Sociological Association (ISA)**

# *Using multilevel models to assess the comparability of examinations*

***John F Bell and Trevor Dexter***

Research and Evaluation Division
University of Cambridge Local Examinations Syndicate
1 Hills Road
Cambridge
CB1 2EU
01223 553849
Fax: 01223 552700
bell.j@ucles.org.uk

**Abstract**

This paper will address some of the conceptual issues that arise when interpreting the results of the multilevel modelling of comparability between examinations. Some of the comparability studies carried out by the Research and Evaluation Division of the University of Cambridge Local Examinations Syndicate will be used to illuminate the conceptual issues involved. The differences in interpretation of the results will be described. The effects of different types of models will be considered.

**Introduction**

In this paper, the objective is to investigate the comparability of different syllabuses for the same subject and for the same time of examination (or a very similar examination). One of the problems that an examination or testing organisation has to face is the issue of comparability. Defining comparability is quite a difficult problem (Bell and Greatorex, 2000). However, in this paper, the studies will involve comparing examinations in the same subject and at the same level which means that the following very strict definition of comparability can be used:

> Two examinations are comparable if pupils who demonstrate the same level of attainment obtain the same grade.

Although a wide variety of methods have been proposed for this type of research, Bell and Greatorex (2000) identified five generic approaches to the problem of investigating this type of comparability:

- Using measures of prior outcomes
- Using measures of concurrent outcomes
- Using measures of subsequent outcomes
- Comparing performance of candidates who have attempted both qualifications *at the same time.*
- Expert judgement of the qualifications

This paper will consider how multilevel models can be applied to the first two approaches. The first three approaches are related and the same methods of analysis can be used to investigate the problem. They have been separated because the advantages and disadvantages are different. The word outcomes has been

deliberately chosen so that it covers the results of a wide range of measures including tests of aptitude, achievement, subsequent job performance.

For methods involving measures of outcomes, the statistical methods used to analyse the data can be the same but the interpretation of the results is different. It is useful to consider what the exact wording of the study results is and how inferences about comparability can be made from each of the first three methods. Assuming that there is no difference between two qualifications, then the results of each study and the assumptions needed to make a valid inference are given in Table 1. These issues have also been considered by Jones (1997).

**Table 1: Results and assumptions of generic approaches to comparing qualifications**

| Method | Strict meaning of results | Some assumptions required for comparability |
|---|---|---|
| Using prior outcomes | The measures of prior outcomes are, on average, the same for the candidates who have obtained both qualifications at a particular level/grade. | If assessing knowledge and skills, then relative/absolute* progress in obtaining them must be the same for both qualifications.<br>If assessing potential, it must be stable over the period between obtaining the prior outcome measure and obtaining the qualification. |
| Using concurrent outcomes | The measures of concurrent outcomes are, on average, the same for candidates who have obtained both qualifications at a particular level or grade. | The attainment is the same only for the skills, knowledge and/or potential assessed by the concurrent measure. The qualifications could differ on other aspects of attainment. |
| Using subsequent outcomes | The measures of subsequent outcomes are, on average, the same for candidates who have obtained both qualifications at a particular level or grade. | There is a causal relationship between achievement required to obtain the qualification and subsequent outcomes.<br>The subsequent outcomes have not been influenced differentially by subsequent events (e.g. the holders of one qualification getting additional training courses). |

*for absolute progress, the measure of prior outcomes produces a score on the same scale as the qualifications.

Comparability studies using a common outcome measure are usually carried out by considering how the relationship between examination performance and the common measure of attainment varies by syllabus (for the purposes of analysis it is sensible to separate syllabuses within boards and possibly options within syllabuses). The data for this type of study has a multilevel structure. The examination candidates are grouped in centres (usually schools or colleges). The leads to the use of multilevel regression models taking the result of the examination as the dependent variable, the measure of prior or concurrent outcomes as one of the explanatory variables and dummy variables for the syllabuses under consideration.

Another issue that arises with the use of multilevel models is that the results of the examinations are expressed as a series of grades. This means that a choice has to be made as to how the grades should be used as a dependent variable. There are three choices. Firstly, the grades could be converted into points and analysing it as a continuous variable using a linear multilevel model. Secondly, the grades can be treated as an ordinal variable and a proportional odds model can used. Finally, a series of binary variables can be created and analysed using logistic regression.

These choices have advantages and disadvantages associated with them. Two examples of the types of statistical comparability study carried out by the Research and Evaluation Division of the University of Cambridge Local Examinations Syndicate. In the next section, the first study using a measure of concurrent attainment and with the grades treated as a continuous variable will be considered. This study considers some of the complexity of fitting a multilevel model. This is followed by an account of a second study that

used a measure of prior attainment and used an ordinal variable and as series of binary variables as response. This section investigate the issues associated with these types of response variables.

**Study 1: Using a continuous variable**

To apply multilevel models to a comparability study, it is necessary to develop an appropriate model. The models in this section are written in terms of a continuous response variable but discussion of the fixed effects hold for any of the responses discussed in this paper (assuming the link function is taken into account).

If there are two examinations and one common measure then the simple regression equation is:

$$y_i = \beta_0 + \beta_1 (com\_meas)_i + \beta_2 (exam)_i + e_i$$

The response variable is a score based on the grade achieved and the explanatory variables are a constant term, the common measure and a dummy variable identifying the examination. If the standards of the two examinations are in line then the term $\beta_2$ is equal to zero and a standard t-test tests this. Incorporating this into a multilevel model, the model becomes:

$$y_{ij} = \beta_{0j} + \beta_1 (com\_meas)_{ij} + \beta_2 (exam)_i + e_{ij}$$
$$\beta_{0j} = \beta_0 + u_j$$

The test for comparability in this case is still a t-test for the term $\beta_2$ just as in the regression case. The standard error for the term in $\beta_2$ in (2) will be slightly higher than in (1). (2) is the better estimate of the standard error, but its larger size means that when multilevel models are used the model is 'correctly' slightly less powerful at detecting significant grading differences.

At this stage it is assumed that the difference between the two examinations is constant across the two examinations. Effectively the two regression lines are parallel. But this is not necessarily the case because, for example, one examination may be easier than another at the bottom of the ability range but equivalent at the top of the ability range. Incorporating an interaction between the common test term and the examination term allows the regression slopes for the two examinations to vary.

$$y_{ij} = \beta_{0j} + \beta_1 (com\_meas)_{ij} + \beta_2 (exam)_{ij} + \beta_3 (com\_meas * exam)_{ij} + e_{ij} \qquad 3$$
$$\beta_{0j} = \beta_0 + u_j$$

The term $\beta_3$ shows the extent to which the sloped varies between the two regression lines. The term $\beta_2$ is now the difference between the two exams where the common measure is equal to zero. It is unlikely to be helpful to see whether exams are comparable at the level of ability equivalent to zero on the common measure, rather it is better to see it for a typical student. Therefore, the common measure needs to be centred on its mean score (or other suitable value) for the $\beta_2$ term to be interpretable. The $\beta_2$ term becomes more interpretable if the common measure is standardised to a mean of zero and a standard deviation of one. The equation becomes:

$$y_{ij} = \beta_{0j} + \beta_1 (st\_com\_meas)_{ij} + \beta_2 (exam)_{ij} + \beta_3 (st\_com\_meas * exam)_{ij} + e_{ij} \qquad 4$$
$$\beta_{0j} = \beta_0 + u_j$$

The difference between the two exams for average candidates is $\beta_2$. For a candidate one standard deviation above average the difference between the two examinations is $\beta_2 + \beta_3$. For a candidate two standard deviation below the average the difference between the two examinations is $\beta_2 - 2\beta_3$.

The model can be made to fit better by including squared terms or other functions. However doing this can decrease the ease of interpretation of this model. Thus it is only worth while fitting additional reference test terms if residual analysis shows the fit for model (4) can be greatly improved upon.

It is often the case that there is a need to control for more than the common measure. Studies have shown that the relationship between examination grade and reference test may vary, for example, by gender or by centre type. If for example girls perform better in an examination than boys of a certain level of the common measure, then the conclusions regarding the comparability of the examinations will be affected by the relative proportion of girls and boys taking the two examinations. This can be controlled for by including gender in the equation. The equation now becomes:

$$y_{ij} = \beta_{0j} + \beta_1(st\_com\_meas)_{ij} + \beta_2(exam)_{ij} + \beta_3(st\_com\_meas*exam)_{ij} + \beta_4(gender)_{ij} + e_{ij} \qquad 5$$
$$\beta_{0j} = \beta_0 + u_j$$

The term $\beta_2$ is the difference between the two exams at the average reference test score after taking gender into account.

This model assumes that the effect of gender is consistent across the two examinations. This is not necessarily the case as it may be that the characteristics of the examinations appeals differently to each gender or it may be that the different social conditions which the exam is taken in means that there are different gender effects across the examinations. In this case an interaction term has to be included between exam type and gender.

$$y_{ij} = \beta_{0j} + \beta_1(st\_com\_meas)_{ij} + \beta_2(exam)_{ij} + \beta_3(st\_com\_meas*exam)_{ij} + \beta_4(gender)_{ij} + \beta_5(exam*gender)_{ij} + e_{ij}$$
$$\beta_{0j} = \beta_0 + u_j \qquad 5$$

There are now effectively four regression lines. This model has a subtle but important difference in interpretation. It now becomes a description of what is happening in the data. We have not taken gender into account but we have shown the effect of gender. It thus becomes interpretative in how we interpret the gender differences in relation to grading standards.

The model may move further away from testing a simple null hypothesis if the situation arises where we no longer expect the regression lines to be co-incident. This can happen when the context of the examinations differ, and it could involve different contextual categorisations between examinations. For example the centre types in one exam may differ to that in another exam which was the situation in the study described later in this section:

$$y_{ij} = \beta_{0j} + \beta_1(st\_com\_meas)_{ij} + \beta_2(exam*ctype1)_{ij} + \beta_3(exam*ctype2)_{ij} + \beta_4(st\_com\_meas*exam*ctype1)_{ij} + \beta_5(st\_com\_meas*exam*ctype2)_{ij} + \beta_6(gender)_{ij} + \beta_7(exam*gender)_{ij} + e_{ij}$$

$$\beta_{0j} = \beta_0 + u_j \qquad 6$$

In this situation there are two possible lines for the first examination to be co-incident on. Thus if centre type is a significant effect and if the exam is coincident on one line it is by definition not co-incident on the other line. It becomes more difficult to interpret. It is up to the researcher and the readership of the research to make a judgement over which centre type the line should match. Or indeed whether the line

should match at all and should lie between the two lines, above the two lines or below the two lines. The process is moving away from testing a single null hypothesis towards a description of the relationships between variables in different groups from which judgements concerning standards are made. Here one shows comparability by showing that the examinations have regression lines behaving how they would be expected to. It involves the researcher or the readership making judgements, which may be challenged.

The applicability of the above models can be demonstrated by considering an example of the use of a concurrent measure of attainment for comparing IGCSE and the GCSE. The IGCSE (International General Certificate of Secondary Education) provided by Cambridge International Examinations (part of UCLES is designed as a two year curriculum programme for the 14-16 age group and is designed for International needs. End of course examinations lead to the award of a certificate that is recognised as an equivalent to the GCSE (General Certificate of Secondary Education) which is designed for UK candidates and is provided by OCR (which is also part of UCLES). Both examinations are single subject examinations and candidates may choose to do several subjects. The examinations are grade in a nine point scale.

The UCLES's Research and Evaluation Division has developed a test of general ability for investigating comparability of a variety of assessments, including those in the several suites of examination provided by UCLES as a whole (Massey and Dexter, 2000). Historically, concurrent measure of outcomes have been extensively used in comparability studies (e.g., Schools Council, 1966; Nuttall, 1971; Willmott, 1977). In more formal psychometric equating, reference tests are referred to as anchor tests. The problem with this approach is that the outcome depends on the relationship between the examinations and the test. A reference test would penalise a board that did not include some of the content of the reference test. This could, of course, be regarded as a valid outcome if it indicated that the examination did not meet a particular subject specification. Christie and Forrest (1981) pointed out that there are three ways of measuring concurrent performance have been used to assess comparability within subjects. Firstly, there are references tests that measure 'general ability', 'aptitude', or 'calibre'. Secondly, there are studies using subject-based reference tests. Finally, a common element can be included as part of all examinations. The UCLES reference test was designed to

The Massey and Dexter study considered fourteen subjects, spanning the full range of the curriculum. These were contrasted with similar GCSE subjects. For the purposes of this paper, results from only mathematics will be considered. For the GCSE examinations, a sample of centres taking OCR syllabuses were selected. These centres were asked to administer the calibration test to all year 11 pupils entering GCSE examinations within two months of the start of the examination session. The pupils were also asked to indicate their gender and whether they normally spoke English at home. The final data set consisted of 3,656 pupils from 43 centres. For the IGCSE, all schools with a 1997 subject entry (the cumulative total of candidate entries for all subjects) of at least 180 from the 30 countries worldwide with the largest IGCSE entries. In this case, 39 schools agreed to take part in the study and returned completed scripts. The IGCSE data set was comprised of 1,664 pupils located in 20 different countries.

The GCSE grades were converted into a points score as follows:

U$\rightarrow$0, G$\rightarrow$1, F$\rightarrow$2, …, B$\rightarrow$6, A$\rightarrow$7, A*$\rightarrow$8.

The following model was fitted

$$y_{ij} = \beta_{0ij} + \beta_{1j}x_{1ij} + \beta_2 x_{2j} + \beta_3 x_{3j} + \beta_4 x_{4ij} + \beta_5 x_{5ij} + \beta_6 x_{6ij} + \beta_7 x_{7ij} + \beta_8 x_{8ij} + \beta_9 x_{9ij} + \beta_{10} x_{10ij} + \beta_{11} x_{11ij} + \beta_{12} x_{12ij} + \beta_{13} x_{13ij} + \beta_{14} x_{14ij} + \beta_{15} x_{15ij}$$

$$\beta_{0ij} = \beta_0 + u_{0j} + e_{0ij}$$

$$\beta_{1j} = \beta_1 + u_{1j}$$

where 0 denotes a constant
1 refers to the standardised calibration test score

2 refers to GCSE comprehensive pupils
3 refers to GCSE independent/selective pupils
4 refers to the interaction between $x_1$ and $x_2$
5 refers to the interaction between $x_1$ and $x_3$
6 refers to IGCSE males
7 refers to IGCSE (comprehensive) males
8 refers to IGCSE (independent/selective) males
4 refers to the interaction between $x_1$ and $x_2$
4 refers to the interaction between $x_1$ and $x_2$
4 refers to the interaction between $x_1$ and $x_2$
4 refers to the interaction between $x_1$ and $x_2$
4 refers to the interaction between $x_1$ and $x_2$
4 refers to the interaction between $x_1$ and $x_2$
9 refers to
10 refers

**Table x: Results of fitting a multilevel model**

Although for the above data, converting to the grade to points and analysing as continuous dependent variable was a reasonable approach. There are circumstances when this approach can be unsatisfactory. This is the case when there are pronounced ceiling and floor effects. This means that the values of the residuals and regressors will be correlated which can result in biased estimates of the regression coefficients (McKelvey and Zavoina, 1975). In addition, Winship and Mare (1984) noted that the advantage of ordinal regression models in accounting for ceiling and floor effects of the dependent variable is most critical when the dependent variable is highly skewed, or when groups defined by different covariate values (e.g. dummy (0,1) variables for each syllabuses) are compared which have widely varying skewness in the dependent variable. This situation does occur in examination databases because some syllabuses attract entries that are, on average, more able than other syllabuses.

**Using an categorical regression models**

It is also possible to consider an examination grade as an ordinal variable (e.g., Fielding, 1999). These can be fitted using the proportional odds model. Although this model solves the problem of floor and ceiling effects by fitting a series of s-shaped logistic curves, there are a number of disadvantages. The proportional odds model is one example of a generalised linear model. There are some problems in using this model. Firstly, parameter estimation in generalised linear models is more complicated than in linear models. This is a particular problem when the multilevel structure is included in the model. However, the main problem with the proportional odds model is not computation but the assumption of identical log-odds ratio for each grade, i.e., the relationship between probability of obtaining a grade and the outcome measure is the same for each grade. Violation of this assumption could lead to the formulation of an incorrect or mis-specified model. In this example, the use of proportional odds model will be considered and the adequacy of the assumptions investigated.

The data used in this example is an approximate 10% sample of centres that entered candidates for GCSE examination taken from a linked database of Key Stage 3 assessment results and GCSE results. There are five different syllabuses under consideration and the sample included 43,366 candidates nested within 460 centres. It should be recognised that the analysis described in this paper was carried out to demonstrate methodology and is not intended to be definitive. Potentially important variables that could have significant implications have not been included in the model. It is for this reason that the examination and the syllabuses have not been identified.

In England (and Wales and Northern Ireland), pupils in state-maintained schools are tested in English, mathematics and science at fourteen years of age (year 9) which is also described as the end of Key Stage 3. These pupils go on take their GCSE examinations in year 11 at the end of Key Stage 4. As a result of the greater interest in school effectiveness and improvement, there has been an increase in the use of value-added measures, this has had lead to the development of linked databases of educational test/examinations results. These databases enable the relative progress of candidates (in most cases, it only possible to measure relative progress and not absolute progress because the two measures will not be on the same scale) to be assessed by various groups of individuals. One problems is that the progress made by different identifiable groups within an entry may have different rates of progress (e.g., Haque and Bell, 2000) and if the composition of the entries of syllabuses vary then it is possible that difference between syllabuses may be exaggerated.

For on ordinal regression, it is useful to assume that there is an unobservable latent variable (y) that is related to the actual response through the "threshold concept" (Hedeker and Gibbons, 1994). The objective of the analysis then becomes a problem of estimating a series of threshold values $\gamma_1, \gamma_2, \ldots, \gamma_{j-1}$, where J equals the number of ordered categories, $\gamma_0 = -\infty$, and $\gamma_J = \infty$. A response occurs in category j (Y=j) if the latent response process exceeds the threshold value $\gamma_{j-1}$, but does not exceed the threshold value $\gamma_j$.

The ordinal multilevel model is defined as follows. Let *i* denote the level-2 units (clusters in the clustered data context, i.e. centres) and let *k* denote the level-1 units (subjects in the clustered data context, i.e. candidates). Assume that there are $i = 1, \ldots, N$ level-2 units and $k = 1, \ldots, n_i$.

Conceptually, the underlying latent variable $y_{ik}$ can be modelled using standard multilevel modelling theory. For a two level model with three independent variables, the model is as follows:

$$y_{ik} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \delta_i + \varepsilon_{ik}$$

where $\beta_0, \ldots, \beta_3$ are estimates for fixed parameters and $\delta_i$, $\varepsilon_{ik}$ are random effects at the centre and candidate level respectively.

With the above the multilevel model for the underlying and unobservable quantity $y_{ik}$, the probability, for a given level-2 unit I, that $Y_K \leq j$ (a response occurs on category j), conditional on the parameter estimates ($\beta_0, \ldots, \beta_3, \sigma_2$), is given by the following equation:

$$\mathrm{P}\left(\mathrm{Y}_k \leq j / \beta_0, \ldots, \beta_3, \sigma_2\right) = \Psi\left(\gamma_j - \mathrm{z}_{ik}\right)$$

where $\Psi\left(\gamma_j - z_{ik}\right)$ is the logistic functions which is defined as follows:

$$\Psi\left(\gamma_j - \mathrm{z}_{ik}\right) = \frac{1}{1 + \exp\left[-\left(\gamma_j - \mathrm{z}_{ik}\right)\right]}$$

and $z_{ik} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \delta_i$. Without loss of generality, the origin and unit of *z* can be chosen arbitrarily. For convenience of interpretation, the origin ($\gamma_0$) is set to zero and the residual is set to $\pi^2/3$.

The above probability is not the most useful. In some contexts, the probability of $Y_K = j$ is of interest and this is calculated:

$$\mathrm{P}\left(\mathrm{Y}_k = j / \beta_0, \ldots, \beta_3, \sigma_2\right) = \Psi\left(\gamma_j - \mathrm{z}_{ik}\right) - \Psi\left(\gamma_{j-1} - \mathrm{z}_{ik}\right)$$

In the context of examinations, the probability of $Y_K \geq j$ (i.e. the probability of obtaining at least a grade C) is of interest. Obviously, this is defined as:

$$\mathrm{P}\left(Y_k \geq j / \beta_0, \ldots, \beta_3, \sigma_2\right) = 1 - \Psi\left(\gamma_{j-1} - z_{ik}\right)$$

The above example assumes that the relationship for each grade is the same shape. This is not necessarily the case. Ways of testing the assumption include investigating the relationship between the variables with an appropriate smoother or by fitting a series of logistic regression models and seeing how the parameters vary. {Cook & Weisberg 1999 #1180} suggest that a smoother may be necessary to visualize the mean function when the response is binary. This technique can also be applied to ordinal data. The use a lowess smoother. It should be noted that it is necessary to consider plots were lines have been fitted for each centre. The plots presented in this paper ignore the multilevel structure. The issue of using plots as preliminary to fitting multilevel models is considered in more detail in Bell (in prep.).

The estimated probabilities for each level of the total Key Stage 3 SAT score for five GCSE English Literature syllabuses have been plotted in Figure 1. Only curves for at least a grade A, at least a grade C and at least a grade F have been plotted. The at least grade A curves are the group at the bottom right of the graph and the 'at least grade F curves' form the group toward the top left. The analysis restricted to these three grade boundaries for two reasons. Firstly, the plot is easier to follow with fewer curves plotted. Secondly, in the process of setting the grade boundaries, only the boundaries for grades A, C and F are set by the awarding committee; the remainder are calculated from these boundaries using a set of rules. It should also be note that there were not many candidates with low Key Stage 3 scores. Historically, English Literature was a subject that was taken by more able pupils and although the entry has increased it stills has this tendency.

It is clear from Figure 2 that there may be some interactions between the total Key Stage 3 score and syllabus. However, it is not clear from this plot was whether the proportional odds assumption holds. It is difficult to be certain that the lines represent curves that have the same shape.

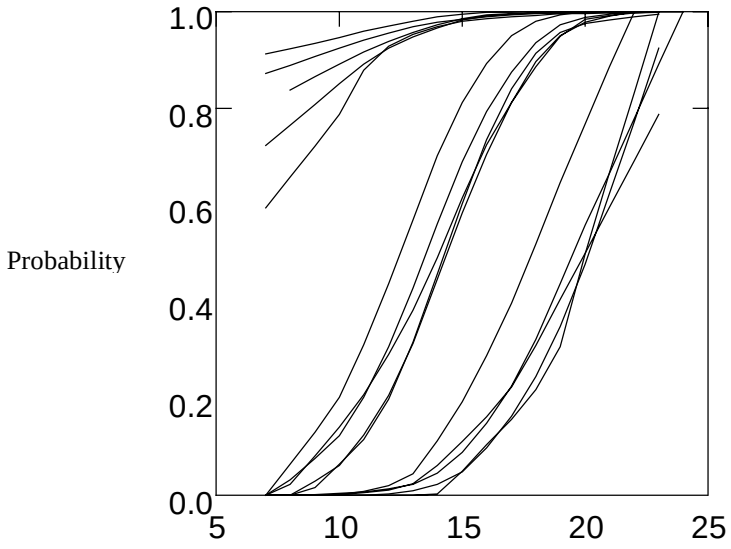

**Figure 2: Lowess smooth curves for each syllabus for at least a grade A, at least a grade C and at least a grade F**

It is possible to investigate the appropriateness of the logit transformations for the estimate mean probablity data used in Figure 2. Obviously if the estimated mean was zero or one, the logit function is undefined. This has been done in Figure 3. Again a lowess smoother has been fitted. Again caution is required when considering the extremes of the line. This plot would suggest that the relationships between total KS3 score and estimated probabilities vary between syllabuses (although it might possible be only one of the syllabus. If this is the case, it would suggest that there is a difference in the structure or marking of syllabuses.

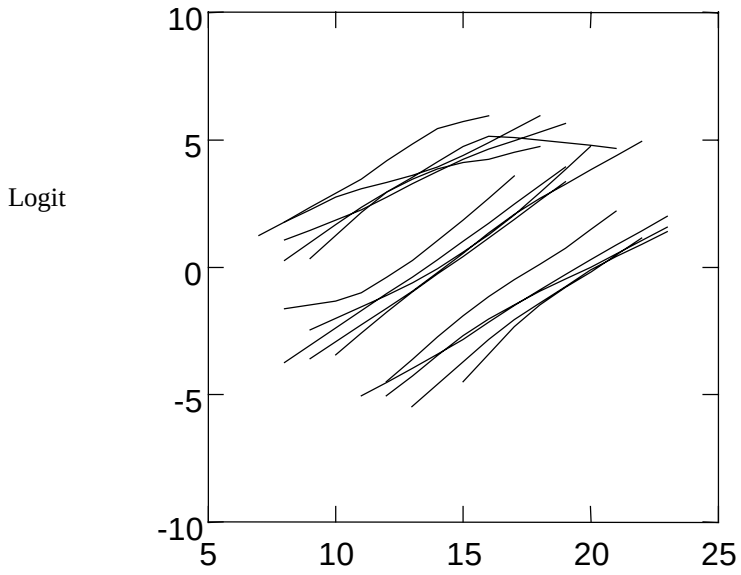


**Figure 8: Logit transformations of the estimated probabilities**

A number of multilevel models were fitted. In Table 2, the results of an ordinal model and a series of binary models are presented. An inspection of the total KS3 parameters and the syllabus parameters indicate that the proportional odds model is not appropriate for this data set because the values vary from grade to grade.

**Table 2: Comparison of proportional odds and logistic regression models**

| Parameter | Ordinal | | At least Grade A | | At least Grade C | | At least Grade F | |
|---|---|---|---|---|---|---|---|---|
| | Est. | s.e. | Est. | s.e. | Est. | s.e. | Est. | s.e. |
| Fixed | | | | | | | | |
| Intercept | -4.42* | 0.08 | -15.16* | 0.19 | -10.86* | 0.13 | -2.54* | 0.24 |
| KS3 total | 0.74* | 0.00 | 0.77* | 0.01 | 0.78* | 0.01 | 0.52* | 0.01 |
| Syll 1 | -0.34* | 0.05 | 0.00 | 0.13 | -0.20 | 0.11 | -0.73* | 0.21 |
| Syll 2 | -0.38* | 0.05 | 0.01 | 0.12 | -0.24* | 0.11 | -0.91* | 0.20 |
| Syll 3 | 0.21* | 0.06 | 0.46* | 0.16 | -0.02 | 0.13 | 0.02 | 0.22 |
| Syll 4 | -0.25* | 0.05 | -0.24 | 0.14 | -0.22* | 0.11 | -0.81* | 0.06 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Random | | | | | | | | |
| Centre | 0.73* | 0.16 | 0.89* | 0.04 | 0.86 | 0.03 | 0.81* | 0.06 |
| Thresholds | | | | | | | | |
| F | 1.09* | 0.04 | | | | | | |
| E | 2.55* | 0.04 | | | | | | |
| D | 4.20* | 0.04 | | | | | | |
| C | 5.84* | 0.04 | | | | | | |
| B | 7.98* | 0.04 | | | | | | |
| A | 9.97* | 0.05 | | | | | | |
| A* | 11.92* | 0.05 | | | | | | |

Although, there is evidence to suggest that the proportional odds model was not appropriate for this set of data, the evidence from the explanatory plots would suggest that the binary models are not satisfactory either. A second series of logistic regression models was fitted. This time with terms representing the interaction between syllabus and Key Stage 3 score. The parameters for these models are given in Table 3. It is clear that the significant interactions only occur for syllabus 1. This is particularly interesting.

**Table 3: Logistic regression models with interaction between syllabus and Key Stage 3 score**

| Parameter | At least Grade A | | At least Grade C | | At least Grade F | |
|---|---|---|---|---|---|---|
| | Est. | s.e. | Est. | s.e. | Est. | s.e. |
| Fixed | | | | | | |
| Intercept | -14.82* | 0.32 | -10.85* | 0.25 | -2.03* | 0.51 |
| KS3 total | 0.75* | 0.02 | 0.78* | 0.02 | 0.48* | 0.04 |
| Syll 1 | -.098* | 0.49 | -0.97* | 0.35 | -2.08* | 0.60 |
| Syll 2 | -0.45 | 0.46 | -0.08 | 0.34 | -1.02 | 0.57 |
| Syll 3 | 1.00 | 0.52 | -0.13 | 0.34 | 0.75 | 0.78 |
| Syll 4 | -1.02 | 0.56 | 0.42 | 0.34 | -1.78 | 0.65 |
| KS3*syll1 | 0.06* | 0.03 | 0.05* | 0.02 | 0.11* | 0.04 |
| KS3*syll2 | 0.03 | 0.03 | -0.02 | 0.02 | 0.01 | 0.04 |
| KS3*syll3 | -0.03 | 0.03 | 0.01 | 0.02 | -0.06 | 0,06 |
| KS3*syll4 | 0.04 | 0.03 | -0.04 | 0.02 | 0.08 | 0.05 |
| Random | | | | | | |
| Centre | 0.90 | 0.03 | 0.86 | 0.04 | 0.81 | 0.06 |

Finally, a third series of logistic regression models were fitted with just one interaction term. The parameters for this series are given in Table 4. There are differences between the syllabuses. Differences in the syllabus parameters could be corrected by changing the boundary marks for the grades but this is not the case when there is an interaction. The difference between syllabus 1 and the other is associated with other measurement characteristics.

One feature of this data set is the difference between grade F and the other two grades. It should be recognised that for this subject the vast majority of pupils achieve at least a grade F. This means that there is not much evidence to make decisions about the boundary when the award is made and the modelling described in this paper is less effective.

**Table 4: Final series of logistic regression analyses**

| | Est. | s.e. | Est. | s.e. | Est. | s.e. |
|---|---|---|---|---|---|---|
| Fixed | | | | | | |
| Intercept | -14.97 | 0.02 | -10.65 | 0.14 | -2.20 | 0.26 |
| KS3 total | 0.76 | 0.01 | 0.76 | 0.01 | 0.49 | 0.01 |
| Syll 1 | -0.82 | 0.42 | -1.18 | 0.29 | -1.91 | 0.42 |
| Syll 2 | 0.01 | -0.12 | -0.24 | 0.11 | -0.92 | 0.20 |
| Syll 3 | 0.47 | 0.16 | -0.01 | 0.12 | 0.01 | 0.22 |
| Syll 4 | -0.24 | 0.14 | -0.21 | 0.11 | -0.80 | 0.21 |
| KS3*syll1 | 0.05 | 0.02 | 0.06 | 0.02 | 0.09 | 0.03 |
| | | | | | | |
| Random | | | | | | |
| Centre | 0.89 | 0.04 | 0.85 | 0.04 | 0.81 | 0.04 |

Using the methodology given in Snijders and Bosker (1999), the explained proportions of the variation for the final series of models were calculated. The results are presented in Table 4. The models do not explain a high proportion of the data in this example data set.

**Table x: Partition of variation for the final series of binary models**

| Proportion of variation: | Grade A | Grade C | Grade F |
|---|---|---|---|
| Explained | 0.54 | 0.55 | 0.37 |
| Unexplained at the centre level | 0.10 | 0.09 | 0.12 |
| Unexplained at the candidate level | 0.36 | 0.36 | 0.51 |

## Conclusions

Using multilevel models has two advantages: it correctly estimates the standards errors of the parameter estimates; and it also provides an estimate of the school level variance. Is this school level variance of importance to the issue of comparability? For the most part not. In the UK the proportion of variance at the school level is fairly low, and so has little impact on the conclusions. However it is possible for the variance at school level to be high in some countries or cultures. When this happens this has implications for the comparability because it may be interpreted as showing that rather than the common measure explaining the performance in the examinations it is the schools the candidates went to which does this. This can challenge the assumptions upon which the definitions of comparability are based upon.

In all the models fitted so far only simple random variation has been allowed. However the model can be fitted with complex random variation. If the complex random variation term is significant then the implications of this for making a judgement on standards will need to be considered.

In the second example, the most important point to note is that using a method based on prior achievement means that it is impossible to determine whether any difference that is found indicates a difference in standard between boards and/or syllabuses or indicates that the candidates made more progress on one syllabus compared with another. Using KS3 data for assessing comparability should be thought of as a screening process to identify potential problems. Other evidence could then be used to investigate the reason for the difference. To be certain of a difference in comparability involves considering the actual attainment of the candidates which means comparing scripts, examination papers and syllabuses.

The use of statistical techniques as described in this paper should be thought of as a screening process to identify potential problems. Except in the case when the concurrent test measures the same thing as the qualifications under consideration, the techniques do not identify whether the attainment is the same. If a difference is found other evidence could be considered to all decisions about the action that could be taken.

There are difficulties in deciding what this action should be. For example, if two syllabuses differ should the standard be changed on one, or both, or at all? One way of investigating the problem further would be to use a study involving expert judgement.

Although they are costly, studies involving judges are very important. Judges have the advantage that they can consider the importance of differences in observed attainment. Studies involving Thurstone paired comparison methodology can be designed to have results that have a high degree of credibility (external judges can be used and if judges disagree this can be identified in the results). Interestingly, it is also possible to use multilevel modelling to investigate data from this type of study but this is beyond the scope of this paper (Bell, in prep.).

The choice between separate binary regression models and the proportional odds model raises an interesting point. Because separate binary regression models are useful for model checking and model building in any case, is there any need for the proportional odds model? The use of separate binary logistic regression models does have some disadvantages. This approach can lead to final models with different sets of covariates for different grades making interpretation difficult (e.g. a sex difference at one grade but not another though this finding is inherently interesting). Categories at the ends of the scale may have very low or very high probabilities, and parameter estimates may not be statistically significant due to less power i.e. the ability of a statistical test to detect differences. The proportional odds model, when it fits, is superior because fewer parameters are fitted. This is better because the standard errors of these parameters are smaller. However, given that any data set considered is likely to be relatively large, the standard errors for the separate binary logistic models are likely to be small (meaning that there is sufficient power to detect small differences between syllabuses). In comparison with the proportional odds models the presentation of the results of binary logistic regressions is much simpler and more interpretable for less experienced users.

## References


Breslow, N.E., and Clayton,D.G. (1993) Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88, 9-25.

Gibbons, R.D., and Hedeker, D. (1997) Random effects probit and logistic regression models for three-level data. *Biometrics*, 53, 1527-1537.

Gibbons, R.D., Hedeker, D., Charles, S.C. and Frisch, P. (1994) A random effects probit model for predicting medical malpractice claims. *Journal of the American Statistical Association,* 89, 427, 760-767.

Goldstein, H. (1991) Nonlinear multilevel models with application to discrete response data. *Biometrika*, 78, 45-51.

Goldstein, H., and Rasbash, J. (1996) Improved approximations for multilevel models with binary responses. *Journal of the Royal Statistical Society*, Ser. A, 159, 505-513.

Haque, Z. and Bell, J.F. (2000*) Evaluating the Performances of Minority Ethnic Pupils in Secondary Schools*. Paper presented at the British Educational Research Conference, University of Wales, Cardiff

Hedeker, D., and Gibbons, R.D. (1994) A random effects ordinal regression model for multilevel analysis. *Biometrics,* 50, 933-944.

Hedeker, D., and Gibbons, R.D. (1996) MIXOR: a computer program for mixed-effects ordinal regression analysis. *Computer Methods and Programs in Biomedicine*, 49, 157-176.

Kim, K.-S. (1990) *Multilevel Data Analysis: a Comparison of Analytical Alternatives.* Ph.D. thesis, University of California, Los Angeles.

Kreft, I. G. G. (1996) *Are multilevel techniques necessary? An overview, including simulation studies.* Los Angeles: California State University.

Littell, R.C., Milliken, G.A., Stroup, W.W., and Wolfinger, R.D. (1996) *SAS System for Mixed Models.* Cary, NC: SAS Institute.

McKelvey, R.D., and Zavoina, W. (1975) A statistical model for analysis of ordinal level dependent variables. *Journal of Mathematical Sociology,* 4, 103-120.

Raudenbush, S.W., Yang, M.-L., and Yosef, M. (1999) Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace approximation. Manuscript under review.

Rodríguez, G., and Goldman, N. (1995) An assessment of estimation procedures for multilevel models with binary responses. Journal of the Royal Statistical Society, Ser. A, 158, 73-89.
Snijders, T. A. B., and Bosker, R. J. (1999) *Mulitlevel Analysis. An introduction to basic and advanced multilevel modeling*. London: Sage Publications.
Tate, R., and Wongbundhit, Y. (1983) Random versus nonrandom coeeficient models for multilevel analysis. *Journal of Educational Statistics*, 8, 103-120.
Winship, C. and Mare, R.D. (1984) Regression models with ordinal variables. *American Sociological Review*, 49, 512-525.
Winship, C., and Mare, R.D. (1984) Regression models with ordinal variables. *American Sociological Review*, 49, 512-525.
Zho, X., Perkins, A.J., and Hui, S.L. (1999) Comparisons of software packages for generalized linear multilevel models. *The American Statistician*, 53, 3, 282-290.