10

MULTILEVEL MODELLING METHODS

Ian Schagen and Dougal Hutchison

Abstract

Aim

The aim of the chapter is to introduce multilevel modelling as a key methodology for the analysis of data in comparability studies and show how it can be applied in different situations and to different data sets.

Definition of comparability

The main definition addressed is that of Cresswell (1996):

> Two examinations have comparable standards if two groups of candidates with the same distributions of ability and prior achievement who attend similar schools with identical entry policies, are taught by equally competent teachers and are equally motivated, receive grades which are identically distributed after studying their respective syllabuses and taking their examinations.

Following the consideration of interaction effects, the chapter suggests limitations with this definition and suggests consideration of a new and more robust definition.

Comparability methods

Although not a comparability method in the same sense as those described in other chapters, multilevel modelling underpins the quantitative approaches discussed elsewhere. It is a statistical modelling tool, derived from multiple regression with the ability to include within-group clustering at a variety of levels in a unified and consistent fashion.

History of use

Since its development in the 1980s, multilevel modelling has been applied in a wide variety of fields, including education, although examination comparability studies have been in some ways a minority application. Over recent years it has tended to replace other less sophisticated analysis methods as the preferred statistical approach. A brief review of studies using multilevel methods is included in the chapter.

Strengths and weaknesses

The main strength of multilevel modelling is its power and flexibility, and ability to model a wide range of scenarios and situations. As with all modelling, the

weaknesses lie in the quality of the available data and problems with setting up models correctly to represent the important underlying relationships.

Conclusion

The main conclusions of the chapter are as follows:

- The advantages of multilevel modelling far outweigh any perceived disadvantages for this kind of work.
- Modelling should explore all possible aspects of comparability, including interactions between boards and key measures such as prior attainment.
- Where such interactions are detected, it is not clear that comparability is maintained – a new definition may be needed to encompass this.

1 Introduction

In this chapter we propose to start from the definition of comparability as given by Cresswell (1996). In this he states that 'two examinations have comparable standards if two groups of candidates with the same distributions of ability and prior achievement who attend similar schools with identical entry policies, are taught by equally competent teachers and are equally motivated, receive grades which are identically distributed after studying their respective syllabuses and taking their examinations.' From this starting point, we aim to show how the use of multilevel modelling techniques can help to investigate comparability understood in this way.

Statistical methods for ensuring comparability may appear to be more objective than those that rely solely on expert judgements, and in many ways this is true. However the objectivity is relative, in the sense that all statistical methods rely on a mathematical model of the underlying situation, and the choice of this model will in most cases affect the results produced and the conclusions reached. There is therefore still an important element of judgement involved in the choice of such models, and in this chapter we aim to inform such judgement with an overview of the range of statistical models available, mostly based on multilevel analysis (see Goldstein, 2003).

In any comparability analysis we are asking questions of the kind: 'What are the differences between these boards/subjects/questions/syllabuses in terms of actual results achieved compared with expected results?' It is in the definition of 'expected results' that the statistical model comes in. The complexity of the statistical model required depends on the data that is available and the assumptions we are able to make about the relationships between examination outcomes and other factors about which we have information, and which may affect examination performance.

At the simplest possible level, we could imagine having no other data than the test scores for two groups of candidates, one of which took Test A and the other Test B. With no further information, our model might be that both groups were equivalent simple random samples from the underlying population and then our statistical test of equivalence would be a two-sample t-test[1]. In this case, the 'expected results' for

the two groups are assumed identical. This equates to the 'no nonsense' definition set out by Cresswell (1996). Obviously this simple assumption is quite likely to be falsified, leading to a lack of robustness and validity in this minimal form of comparability study.

Moving to a more complex and perhaps more reasonable example, let us assume we still have the two groups doing different examinations, but in this case we have a great deal more background information on the two groups, including a number of measures of prior attainment in earlier tests, background information such as the candidates' sex, ethnicity and social status (perhaps even parental income), as well as data about the institutions in which they are studying.[2] We now have much more scope for computing 'expected results', based on a complex regression model taking account of all these factors. We need to be aware, however, that decisions about which variables to include in a comparability study model are a matter for judgement, not just a technical issue depending only on the information that happens to be available.

Once we have reached agreement on which factors should be controlled for, there are some extra complications we would want to take into account.

- Candidates are grouped into institutions or examination centres – probably there is more similarity in outcomes between candidates in the same centre than between centres. Also, relationships with (for example) prior attainment may vary from centre to centre – the so-called 'random slopes' situation.
- There may be interactions between results for the two different tests and background factors. For example, Test A may produce better results for boys rather than girls relative to Test B, or one test may have a stronger relationship with prior attainment than the other. If this is the case, of course, it raises a number of issues about comparability and whether different 'adjustments' should be made for different groups of candidates to bring the tests into line.

Both the above complications can be taken account of by using a suitably complex model, with a structure that allows explicitly for these inter-relationships. The use of multilevel modelling, with which this chapter is largely concerned, will help us to deal with the first complication above. The second complication can also be dealt with in the setting up of the model by including suitable interaction terms. The identification of such potential relationships and the inclusion of them in the model used are very important elements of any comparability study, and will form a major part of the theme of this chapter. However, modelling does not solve the interaction problem, it merely allows us to quantify it. It could be argued that, by the Cresswell definition, as soon as statistically significant interaction terms are detected then comparability is violated – since we could find a sub-set of candidates with identical characteristics but different results in the two examinations.

Throughout this book there are issues that need to be addressed in the course of any comparability study, and these will not all be rehearsed here. An example, however,

is the issue of unmeasured factors that are confounded with the differences we are interested in, the outcomes from the different examinations.

Let us suppose that the Test B syllabus is more attractive to candidates and encourages a more positive motivation and response to the subject, and hence a better set of results. In the model we set up, we cannot control for this and must assume that motivation and response are the same across the two groups. The comparability study will therefore adjust the results of the unmotivated Test A group to be equivalent to those of the motivated Test B group. Is this fair? If we were able to measure motivation and allow for it in the modelling, then Group B would be acknowledged to have achieved comparatively better results than Group A, and as the study has not shown this it has actually failed to achieve true comparability between the *examinations* rather than the syllabuses. For further discussion, see Jones (1997).

It is possible to think of other examples where unmeasured confounding factors or selection effects can lead to misleading results. The only rigorous way in which such confounding effects can be eliminated is through the adoption of a Randomised Control Trial (RCT) approach (see Mosteller & Boruch, 2002; Styles, 2006). In this approach, candidates or centres would be randomly allocated to syllabus A or B and thus would take the equivalent Test A or Test B. Because of the randomisation, confounding factors would be equally likely to apply to either test, and if sufficiently large samples were used it should be possible to carry out a powerful test for the comparability of the two syllabuses. (Note, however, that this would not overcome the difficulty set out in the previous paragraph – differential motivation between syllabuses. This could only be detected by random allocations to examinations as well as syllabuses, with further administrative and practical difficulties.)

A number of practical and ethical objections could be raised to this particular design of comparability study. One is that it would be an administrative and logistical nightmare to assign candidates within the same centre to different syllabuses, and that only randomisation at the centre level would be at all feasible. This modification of the RCT design would work also, but would suffer from the problems of correlations within centres and would thus require a larger sample size to detect a given difference. Although in theory an RCT 'randomises away' all the effects of related variables, and could therefore be analysed by a simple t-test, in practice the use of a multilevel model even in this case would be recommended, for two reasons. One is that it would allow for the effects of measured background factors that were not completely balanced between the two groups. The other is that it would allow for the within-centre clustering mentioned above, if randomisation occurred at the centre level.

Of course, this kind of comparability study never happens, for what may well be good reasons to do with practicality, customer choice and other pragmatic considerations. Without going into the arguments in favour of attempting such a study, for our purposes we shall treat it as an ideal and see to what extent the examples we shall consider in this chapter fall short of this design. In the meantime,

let us just list some of the effects that can and cannot be taken into account when modelling administrative data rather than analysing a full RCT.

The following can be included in a suitable model:

- overall effects of measured background factors on performance, plus non-linearities in these effects
- interactions between measured factors in their effect on performance
- clustering of candidates within centres
- random variations between centres in the effects of background factors
- interactions between different examinations and background factors.

The following cannot generally be allowed for in modelling:

- options effects – candidates or centres preferentially choosing different examinations[3]
- other unmeasured factors, especially those that vary systematically across syllabuses
- differences between examinations that mean they are testing different constructs, or mixtures of constructs.

Sophistication of modelling does not guarantee the validity of a comparability study – on the other hand, unsophisticated models may be missing something critical that fatally challenges their validity. In this chapter we shall set out some of the features of complex models and the advantages they can bring, but it will always be important to bear in mind the caveats and health warnings expressed above.

2 Advantages of using multilevel modelling

A widely used technique in statistics and in research with educational applications is regression. This explores how a number of variables, described as explanatory variables, relate to another variable, referred to as a response or outcome variable. Explanatory variables are also sometimes referred to as independent variables or predictor variables, and outcome variables are also referred to as dependent variables. An example would be to predict the score on a later test given knowledge of earlier test scores.

The earliest and probably still the best known type of regression is known as Ordinary Least Squares (OLS) regression. In this the outcome Y is assumed to be some function (often a linear function) of the explanatory variables X, Z; and to take account of the fact that one does not expect such a relation to be exact, an error or residual term e is introduced. To distinguish the cases, each one is numbered, using

the suffix *i*, so that we get Y_i, X_i, Z_i and so on. Then a very simple relationship with one X variable can be written in equation form:

$$Y_i = \beta_0 + \beta_1 X_i + e_i \qquad (1)$$

β_0 is described as the *intercept* and β_1 as the *slope*.

To provide a worked example to illustrate the differences between OLS regression and multilevel modelling and to show the kinds of analyses that are possible, we have constructed simulated data with an assumed underlying structure, as follows:

- Twenty candidates in each of twenty centres have GCSE subject results represented as Uniform Mark Scale[4] (UMS) scores (Y) plus a measure of prior attainment based on Key Stage 3 fine grades[5] in the same subject (X).
- There is a linear relationship between X and Y, plus a random error quantity for each candidate.
- Each centre is classified as either Type A or B; the relationship between X and Y is in general slightly different for the two types of centre.
- In addition, the relationship between X and Y varies from centre to centre.

The explanatory variable is the Key Stage 3 fine grade (X) and the response variable is the UMS score (Y). Equation (1) can have values of its coefficients (β_0 and β_1) estimated by a standard OLS regression package from the simulated data set; in this case the estimates are shown below:

$$Y_i = -106.93(18.44) + 79.82(3.05) * X_i + e_i \qquad (2)$$

This shows that, on average, for an increase of one point in X, there is a corresponding increase of approximately 80 points in Y. The figures in brackets are the standard errors of the coefficients. These indicate the uncertainty in the estimates due to the fact that they are derived from finite data sets. Assuming the error terms in the model are Normally distributed (see later), then from the standard error estimates it is possible to derive 'confidence intervals' for the coefficients, such that there is a specified chance that the true value of each coefficient lies within its given interval. For example, the coefficient of X has a standard error (SE) of 3.05 – to derive the 95% confidence interval for this value we multiply by 1.96 and add and subtract this value from the estimate. This yields an interval from 73.84 to 85.80. The constant term (–106.93) is the intercept, the expected value of Y if X were to have the value zero – in practice, not ever attained but a necessary element of the model.

The e_i error term in this model is assumed to be independent of the response variable and the explanatory variables, and to be normally distributed with mean zero. Explanatory variables do not have to be continuous, so we could include a categorical variable – for example, to compare Type A and B centres by giving Z a value of 0 for Type A and 1 for Type B.

The estimated coefficients corresponding to this extended model are given by

$$Y_i = -106.38(18.66) + 79.83(3.05) * X_i - 1.17(5.88) * Z_i + e_i \qquad (3)$$

The 95% confidence interval for the difference between Type A and B centres (taking account of prior attainment) is –12.69 to 10.35, implying there is no clear evidence from this analysis of a real difference overall between the two centre types.

This kind of model makes a key assumption that the candidates in the analysis are all equally representative of candidates in general, and takes no account of the fact that they are grouped within centres. It is frequently the case, however, that candidates in the same centre are more similar than they are to candidates in other centres.[6] This means that it is not legitimate to use the standard OLS regression, which assumes that all units are independent. We could get biased results, in particular for the standard errors in the coefficients, which could be underestimated by assuming that all the observations were independent.

The OLS model (1) above assumes that there is no additional information obtained by knowing the higher-level unit (centre) from which a lower-level unit (candidate) comes. One possible approach to taking account of this would be to define a set of 'centre effects', one per centre, and include these in the OLS model. This obviously makes the model much more cumbersome, and also assumes that the centre effects are to be treated as *fixed* – in other words, we are interested in these values in their own right, rather than as a general addition to the uncertainty in the modelling. If we turn to the multilevel modelling approach, these centre effects are treated as *random* – we are only interested in their overall effects and the differences they make to the model as a whole. We shall now consider this approach to the analysis of the same data. Model (1) can be extended by including a term to take account of the similarity of items within higher-level units.

$$\begin{aligned} &Y_{ij} = \beta_0 + \beta_1 X_{ij} + u_j + e_{ij} \\ &u_j \sim N(0, \sigma_u^2),\ e_{ij} \sim N(0, \sigma^2) \end{aligned} \qquad (4)$$

The term u_j ('centre residual') is assumed to include all those unmeasured factors at a centre level that influence results for all candidates at the centre; in general we assume this combined factor is Normally distributed. However, in certain cases such as modelling binary outcomes (see later) a non-linear model may be required, although centre residuals may still be assumed to be Normally distributed in the transformed metric.

Each equation now has two subscripts, *i*, *j*, corresponding to candidate and centre. This means that there is a separate regression for each higher-level unit. In this equation, because β_1 has no *j* subscript, the relationship between *Y* and *X* is the same in each centre, and these regressions are all parallel. This is an example of multilevel modelling.

Estimating the same model as in (3) above, but taking account of within-school similarities using multilevel modelling, gives:

$$Y_{ij} = -101.65(18.89) + 79.03(2.81) * X_{ij} - 1.15(12.43) * Z_j + u_j + e_{ij} \quad (5)$$

There are small differences between the coefficients in the two sets of results, but the main difference lies in the standard error of the *Z* coefficient, which has increased from 5.88 under the OLS estimation to 12.43 under multilevel modelling. This is a fairly common feature of moving from OLS regression to multilevel modelling: standard errors for variables that relate to higher-level units tend to increase, due to the clustering of data within such units. In the above example it makes no difference to the significance of the coefficient, but it is easily possible to find cases in which this difference can affect the conclusions drawn from the analysis.

When setting up our example data we said that the relationship between *Y* and *X* was different for the two centre types, and also that it varied from centre to centre. To see how both these effects can be included in the modelling we will discuss interactions and random slopes. We have seen that there is no apparent significant difference between the centre types when we assume the regression lines are parallel; however, to model a non-parallel situation we need to define an interaction term:

$$I_{ij} = Z_j (X_{ij} - 6) \quad (6)$$

For Type A centres, the value above is zero; for Type B it introduces a change in the regression slope against prior attainment. The value 6 is the mean value of *X* and ensures that the interaction term is zero on average or 'centred'. Including this extra term in the model gives us the following fitted model:

$$Y_{ij} = -25.82(23.85) + 66.31(3.73) * X_{ij} - 0.39(12.18) * Z_j + 27.16(5.45) * I_{ij} + u_j + e_{ij} \quad (7)$$

The coefficient of the interaction term is clearly significant, and this implies that the two centre types do have different regression slopes: 66.31 for Type A and 93.47 for Type B. In order to look in more detail at the above model, we need to consider the so-called 'random part' of the model – the variances and covariances between the various parameters which vary from candidate to candidate and from centre to centre. In the above model there are only two elements to this, and the estimated variances and standard errors are set out below:

Between-candidate variance:	2629.5 (190.8)
Between-centre variance (intercept):	609.7 (234.6)

From the above figures we can surmise that variation between centres accounts for about 19% of the total variance in the outcome, once other factors are allowed for.

To add an extra complexity to the above model, let us assume that the regression slopes vary from centre to centre, as well as between centre types. To model this we assume that the coefficient of *X* is made of two parts:

$$\beta_j = \beta_1 + u_{1j} \tag{8}$$

where the first term is the overall fixed part of the coefficient and the second, with mean zero, is the part that varies from centre to centre. The centre-level covariance matrix becomes:

$$\begin{bmatrix} Var(u_{0j}) & Covar(u_{0j}, u_{1j}) \\ Covar(u_{0j}, u_{1j}) & Var(u_{1j}) \end{bmatrix}$$

Fitting this model to the data gives us the following:

$$Y_{ij} = -27.95(26.75) + \{66.64(4.68) + u_{1j}\} * X_{ij} - 0.137(12.48) * Z_j + 26.07(6.74) * I_{ij} + u_{0j} + e_{ij} \tag{9}$$

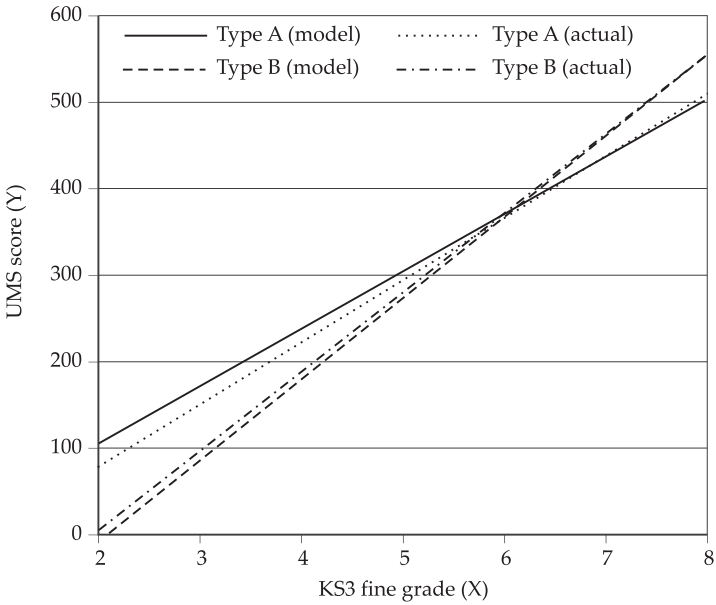
In order to look in more detail at the above model the random parameters are set out below, together with the standard errors of the estimates:

Between-candidate variance:	$Var(e_{ij})$:	2551.9 (189.9)
Between-centre variance (intercept):	$Var(u_{0j})$:	2053.4 (2318.6)
Between-centre variance (slope):	$Var(u_{1j})$:	78.8 (70.8)
Centre covariance (intercept/slope):	$Covar(u_{0j}, u_{1j})$:	–353.3 (395.3)

From the above, it seems that the standard errors in the slope variance and the intercept/slope covariance are both close to the estimated values, implying that neither is statistically significant. To confirm this more rigorously, we should consider the change in the likelihood (represented in MLwiN by the 'deviance') in adding the random slopes. This gives a change of 3.63, compared with a critical chi-squared value (5% significance, two degrees of freedom) equal to 5.99. From this, it seems that this extension to the model does not result in significant random parameters at the centre level – in particular the assumption of random slopes is not supported by the data. It should be noted that this does not mean there are no random slopes, but rather that our data is insufficient to detect them with confidence.

Reverting to model (7) above, Figure 1 shows a graph of expected UMS score (*Y*) versus prior attainment (*X*) for each type of centre. It shows the lines based on the model results above, plus the 'actual' relationships on which the original simulated data was based.

Figure 1 Expected outcomes versus prior attainment by centre type (model and actual)



It is clear that we have made a reasonable job of recreating the underlying relationships from a relatively small amount of data. However, normally we do not have the luxury of knowing what these relationships should be, and our model fitting process cannot guarantee that we have uncovered all possible details of the structure of the data. Models of increasing complexity, with additional terms, interactions, and random parameters can be created, and this kind of model fitting relies heavily on the modeller's insights and judgement rather than explicit rules. There is no substitute for experience and a deep understanding of the subject matter when engaged in this kind of activity. In this section we have used simulated data to illustrate some aspects of the different kinds of models that are possible; in later sections we will be using real data to illustrate particular applications in examination comparability studies.

What difference does it make if we do not use multilevel modelling? Some authorities argue that it is not necessary to use multilevel modelling if the within-school clustering is sufficiently small. This can be assessed by comparing the variance at the higher level with that at the lower level; however, in order to assess this it is necessary to run a multilevel model or equivalent. Even where this holds and the differences between intercepts are negligible, there is still the possibility that there will be differences between slopes, so it is advisable at least to investigate the use of multilevel modelling.

Other approaches include, on the one hand, ignoring the level structure and simply attaching all variables, no matter what level they arise from, to the lowest-level unit (in this instance, pupils); or, on the other hand, aggregating lower-level scores, and

dealing with aggregates (here, schools). The first of these risks giving a highly exaggerated impression of the effect of higher-level effects. The second is effectively throwing away data, and largely ignores individual-level pupil differences and differences in slopes between higher-level units. Also, using this approach means that the ecological fallacy[7] can give a completely biased and inflated estimate of individual correlations.

Another approach, which goes some way towards a multilevel approach, carries out a series of separate OLS regressions, one for each higher-level unit, and then attempts to model the residuals from each (Burstein, 1980). However, this can give biased estimation and also loses sight of the essential unity of the data. It can also lead to excessive numbers of parameters and loss of parsimony, as well as giving no way to generalise to the population of higher-level units. Multilevel modelling, by treating every effect at the appropriate level, gives unbiased estimates of standard errors, and enables the modelling of between-level interactions. Multilevel modelling, generally, provides a unified treatment for effects at all levels. It is efficient in terms of the number of parameters to be estimated and allows the extension of existing generalised linear model techniques by taking account of hierarchical structures in the data.

The value of using multilevel modelling has been attacked by some writers (for example Fitz-Gibbon & Tymms, 2002; Gorard, 2003a). While in general they appear to accept in principle the benefits that may be adduced by using multilevel analysis, they are unhappy with the widespread use of it on two main grounds: first, that it is complicated to understand the details, thus potentially alienating users and audience; second, while apparently implicitly accepting that multilevel modelling is a technically superior exercise, they argue that it fails to produce any new results, and that the results are closely correlated with those from Ordinary Least Squares. Gorard has a number of other theoretical points and these may be assessed in the debate between Gorard and Fielding & Plewis (Gorard, 2003a, 2003b; Plewis & Fielding, 2003).

The writers of this chapter favour the use of multilevel modelling methods in these applications and more generally. First, we believe that the objections to the difficulty of the technique are overstated, and that it is more important that analyses such as these are carried out to the best of our available techniques – it is more important to be correct than to appear simple. Second, the main difference arising as a result of using multilevel methods is not that new real findings are made that would be otherwise overlooked, but rather that we avoid making findings that are not there. That said, it is often true that the values of fixed effects found in these analyses will be similar to those found using OLS regression, although this is not true in all cases.

3 What can be compared using multilevel modelling?

There are many ways in which multilevel modelling techniques (Goldstein, 2003; Raudenbush & Bryk, 2002) may be used in examination comparability studies. Very simply, their strengths may be summed up as:

- they are a type of regression technique used to compare like with like
- they take account of the structure of the education and examination system, and allow for the fact that candidates within entry centres or teacher groups are likely to be more similar to each other than to the rest of the population as a whole.

We shall describe the statistical basis of a number of applications. Nothing is said here at this stage about the realism or otherwise of any assumptions used. Later, we shall also describe applications to real data, both from our own work and that of other researchers, and comment on some of the features and assumptions. For this section of the description, a simple linear model treating the outcome examination result as a continuous variable is used. Later in this chapter we show how different types of model can give different results.

3.1 Application 1: Comparing different boards for the same subjects

The characteristics of the pupils taking their exams via different boards may well be very different, so a number of proposed factors are included in the analysis. The equation is given by

$$Y_{ijk} = \beta_0 + (\sum_p \beta_p x_{pijk}) + v_k + u_{jk} + (\sum_q \beta_q z_{qijk}) + (\sum_q e_{qijk} z_{qijk}) \qquad (10)$$

where:

Y_{ijk} = the examination outcomes (assumed to be a continuous variable, e.g. UMS score) for candidate *i* in teaching group *j* in centre *k*

β_0 = the intercept (the expected value of Y_{ijk} when all variables are equal to zero)

$\sum_p \beta_p x_{pijk}$ = the sum of the coefficients for the explanatory variables times the value of the variables for candidate *i* in teaching group *j* in centre *k*. Explanatory variables may include prior attainment, sex, age and other relevant background characteristics.

z_{qijk} is an indicator variable for the board *q* (= 1 if candidate sits board *q*, = 0 else)

β_q is the coefficient for board *q* (the amount by which the expected scores for board *q* differ from β_0, when all variables except board indicators are equal to zero)

v_k = the effect of centre *k*, assumed Normally distributed with mean zero

u_{jk} = the effect of teaching group *j* in centre *k*, assumed Normally distributed with mean zero

e_{qijk} = residual error for candidate *i* in teaching group *j* in centre *k*, at board *q*, assumed Normally distributed with mean zero and variance σ_q^2. This allows for a different variance for each board.

In setting up such models, there are two kinds of background variables that can be included – those which are assumed to be numerical scales (such as previous test scores), and those which are categorical (such as board taken). To include categorical variables, we produce a set of binary indicators (taking values 0 or 1) to represent each category *except one*. The omitted category is the 'default' or 'base' category against which the others are tested. It is important not to include indicators for all categories, otherwise the model becomes 'over determined' and fails to run.

If the coefficients of the β_q are statistically significant, then the results for the boards are considered as different in overall level.

It is not just the overall levels of attainment that are of interest. The comparative spread of grades within the boards should also be considered. Thus a board could give 'too many' (however defined) grade A passes, but compensate for this by giving give 'too many' (however defined) grade F, so that while the distribution of grades was quite different, the mean levels were the same. This can be investigated by comparing the within-board variances. If there is a statistically significant difference between the estimated values of σ_q^2 for the different boards, that is, for different values of *q*, then the spread of grades can differ between boards. This is an example of 'complex variance modelling', which will be discussed in more detail later in the chapter (see section 6).

A variation of the above application is comparing 'standard' and alternative syllabuses within boards for the same subject. The approach and equations are the same as those given above, except that the values of *q* relate to the two syllabuses.

So far this treatment has dealt with the scenario where the groups of candidates taking each type of examination are distinct, and we had to attempt to equate these by taking account of other measured characteristics of the individual. An alternative is where each candidate takes more than one exam, and thus performance may be compared more directly (for example, in a variant of the above, some candidates may be entered for more than one board). This is dealt with next.

3.2 Application 2: Comparing results from different boards for the same subjects taken by the same or overlapping sets of candidates

If some or all of the candidates in the study have taken the two examinations to be compared, then the analysis is essentially multivariate (see Goldstein, 2003, pp. 139ff). In this situation an additional lowest level (board within candidate) may be proposed. If we are dealing with a candidate within centre model, and teaching group is disregarded, then the equation for two boards is given by a three-level model:

$$\begin{aligned} Y_{ijk} &= \beta_{1ijk} board_1 + \beta_{2ijk} board_2 \\ \beta_{1ijk} &= \beta_1 + v_{1k} + u_{1jk} \\ \beta_{2ijk} &= \beta_2 + v_{2k} + u_{2jk} \end{aligned} \tag{11}$$

where:

β_1 = the grand mean for board 1

β_2 = the grand mean for board 2

v_{1k} = the effect of centre *k* on board 1

v_{2k} = the effect of centre *k* on board 2

u_{1jk} = the level 2 effect of candidate *j* in centre *k* on board 1

u_{2jk} = the level 2 effect of candidate *j* in centre *k* on board 2.

There is no level 1 variation because of the way the problem has been set up (see section 4). Because the exams are being taken by the same candidates, and generally in the same centre, they are assumed to be correlated:

$$\begin{bmatrix} v_{1k} \\ v_{2k} \end{bmatrix} \sim N(0, \Omega_v): \ \Omega_v = \begin{bmatrix} \sigma_{v1}^2 & \sigma_{v12} \\ \sigma_{v12} & \sigma_{v2}^2 \end{bmatrix}$$

$$\begin{bmatrix} u_{1jk} \\ u_{2jk} \end{bmatrix} \sim N(0, \Omega_u): \ \Omega_u = \begin{bmatrix} \sigma_{u1}^2 & \sigma_{u12} \\ \sigma_{u12} & \sigma_{u2}^2 \end{bmatrix} \tag{12}$$

The covariances indicate a common factor for that board. As such, they are comparable to an overall within- or between-centres variance for that subject, but with fixed board-specific factors taken into account.

If the difference $\beta_1 - \beta_2$ is statistically significant, then the results for the boards are considered as different in overall level. Similarly, the values of the within-board variance may be compared at centre level, or at candidate level; or the sum of the two levels may be compared.

If the assumptions of the model are met, then this is a more powerful model than the two-sample model considered previously. In principle, it is not necessary to include any more home background or prior attainment predictor variables, since all differences between individuals are allowed for in a more powerful fashion than by attempting to measure them. Other variables may be included, if it is considered that the contrasts vary with other factors. On the other hand, variables involved in the pupils' experience of taking exams, such as number of hours studied per week or motivation, could be included if available.

This approach will give an average difference between boards for the pupils involved. For this to be an unbiased estimate of the overall difference between the boards the candidates would have to be a random sample of all candidates. In fact this is unlikely, since candidates taking the same subject twice, with different boards, are likely to be untypical. On a practical level, timetabling issues can prevent candidates taking the same subjects with different boards under live examination conditions. This application relates strongly to Chapter 9, on common examinee methods.

3.3 Application 3: Comparing optional questions within an examination paper

In a sense all questions within an examination paper are optional, to the extent that pupils may choose to do any sub-set of them. In this situation, a possible multilevel modelling approach could be similar to that used for comparing boards. The model could be a three-level one, for example questions within candidates within centres. In this model questions are treated as fixed effects, as we have a separate model parameter that is estimated for each question. However, candidates' responses to these questions form the bottom-most level and include a random element. If there were any further or intermediate levels, for example teaching unit within centre, then a four-level or higher model might be used. This kind of modelling assumes that the omitted questions form a random pattern (see discussion in Yang *et al.*, 2002).

This illustrates the fact that multilevel models can be used in a whole range of applications, provided it is possible to cast the situation into terms that can be analysed in this way. As experience and fluency with manipulating these kinds of model grows, the analyst will be able to see ways in which more complex situations become amenable to analysis in this way. Goldstein (2003) contains a range of different applications of multilevel modelling to social and educational data.

So far we have said nothing about the realism or otherwise of these models and their inherent assumptions. Later in this chapter, we give some examples of the application of these models, both from our own research and those of other workers, and discuss aspects of the modelling assumptions.

4 Examples of modelling different structures

As we start to consider the use of multilevel modelling in comparability studies, we need to model two quite different possible structures in the data, depending on the procedures used in the particular study. There are essentially two data structures that we need to consider:

1. **Separate forms**: Each individual in the study completes just one form of assessment (one board, or syllabus or subject, etc.). Comparability between forms is therefore evaluated through relationships between outcomes and other common background factors. These are exemplified in Application 1 in section 3.

2. **Multiple forms**: Each individual in the study completes more than one of the forms of assessment being compared, and comparability is evaluated in a more 'direct' fashion. These are exemplified in Applications 2 and 3 in section 3.

In this section we will consider how to use multilevel modelling to analyse both structures, with worked examples.

4.1 Example 1: Separate forms

As an example for this we shall take data from the study into alternate forms of the GCSE mathematics examination carried out in 2005 (see Stobart *et al.*, 2005). In this case we shall not consider the alternatives to the existing three-tier structure, but use data supplied by four different examining boards on results in their three-tier version of the examination. In addition to the GCSE results for these candidates, information was available on the centre in which they entered and their Key Stage 3 (KS3) results, in terms of 'fine grades' in all three core subjects.

Results for the three-tier mathematics examination were presented for each candidate in terms of both grade awarded and Uniform Mark Scale (UMS) score. For this example, the UMS scores were harmonised to give 60 points per grade (e.g. grade A = 540 to 600; grade B = 480 to 539; etc.), and these were used as outcomes for the modelling, on the assumption that they could be treated as essentially numerical outcomes rather than categorical.

Table 1 shows the basic statistics for the UMS scores for each board, out of the total of 7,347 cases with complete data.

In the model set up to analyse this data we assume a single outcome (UMS score) but a number of background factors that may affect the outcome. One group relates to prior attainment, and includes the KS3 mathematics fine grade result for each candidate, plus (possibly) their KS3 results in English and science. The assumption here is that, given candidates with equal KS3 results their GCSE outcomes should be, on average, equivalent irrespective of the board taken. In order to test this assumption we need to include in the model indicators related to the board taken.

Table 1 GCSE mathematics UMS scores by board

Board	Mean UMS score	Standard deviation	Number of cases
Board A	389.9	102.4	3147
Board B	337.1	56.5	1330
Board C	378.8	93.3	858
Board D	388.5	99.8	2012
Total	**378.7**	**95.9**	**7347**

With this in mind, the variables included in our Model 1 for this example are:

- **Cons** – a constant term (= 1) whose coefficient represents the intercept on the vertical axis when all factors are set to zero
- **Board B, Board C, Board D** – indicators for three of the boards relative to Board A, the 'default'
- **KS3mfine, KS3efine and KS3sfine** – fine grade measures in maths, English and science.

The outcome for this model was **T3umstot**, the UMS total score for the three-tier examination. Before putting the data into the multilevel modelling, an OLS regression was run using SPSS with the results shown in Table 2.

Table 2 OLS regression coefficients for Example 1 Model 1

	Variables	Estimates from modelling		
Name	**Description**	**Coefficient**	**SE+**	**Significant?**
T3umstot	*UMS total score*	*Outcome variable*		
Cons	Constant term	–136.90	3.99	*
Board B	Board B indicator (vs. A)	25.72	1.62	*
Board D	Board D indicator (vs. A)	18.53	1.36	*
KS3mfine	KS3 maths fine grade	57.26	0.94	*
KS3efine	KS3 English fine grade	7.65	0.85	*
KS3sfine	KS3 science fine grade	18.85	1.09	*
Board C	Board C indicator (vs. A)			Omitted – not significant

* = significant at 5% level

+ Here and in other tables 'S.E.' is the standard error of the estimate in the preceding column

The results of this modelling imply that there is no statistically significant difference between Boards C and A in terms of results controlling for KS3 attainment, but that both Boards B and D seem to produce higher UMS scores than would have been predicted for Board A. However, the OLS model takes no account of within-centre clustering of candidates and the next step is to turn to multilevel modelling to deal with this. The same basic model was run using MLwiN (Rasbash *et al.*, 2005), and the results are shown in the equations window from that program reproduced as Figure 2, as well as in Table 3.

Figure 2 MLwiN output for Example 1 Model 1

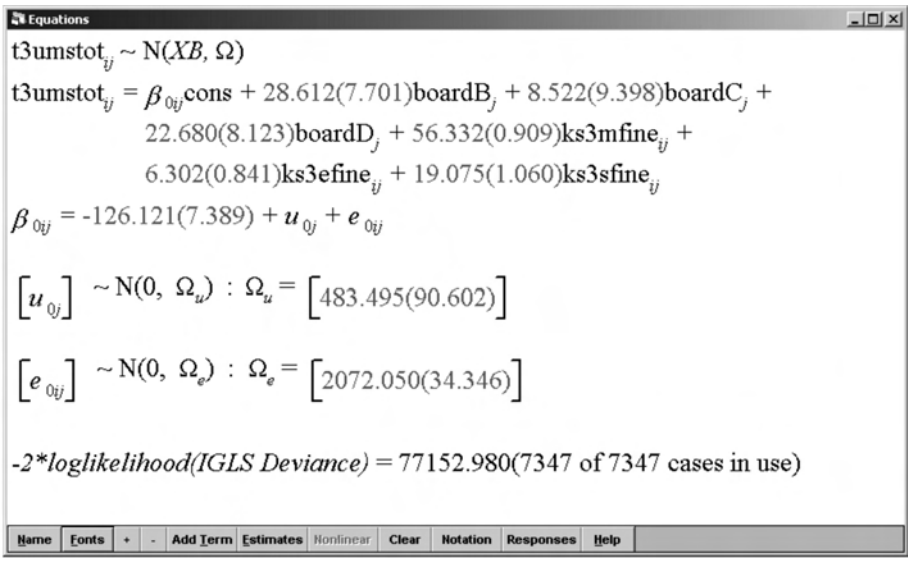


Figure 2 shows the model fitted to the outcome (variable name 't3umstot'), with estimated coefficients and standard errors. The centre- and candidate-level error terms are also included, and the random variance estimates of these are also given. The final line is the 'deviance' (–2*log likelihood) and gives a measure of the extent to which the model explains the data, which can be compared with the same measure for alternative models. These model parameter estimates are replicated in Table 3.

In many ways, the results are very similar to those obtained from the OLS analysis – coefficients are similar in magnitude, and all are clearly significant except for the Board C effect. However, if we look at the coefficient standard errors (in brackets in the MLwiN output) we can see some clear differences. The standard errors for the KS3 fine grades are actually very similar, but for the board variables there are real differences. The standard errors from the multilevel modelling analysis are

Table 3 Multilevel modelling coefficients for Example 1 Model 1

Variables		Estimates from modelling		
Name	**Description**	**Coefficient**	**SE**	**Significant?**
T3umstot	***UMS total score***	***Outcome variable***		
Cons	Constant term	–126.12	7.39	*
Board B	Board B indicator (vs. A)	28.61	7.70	*
Board C	Board C indicator (vs. A)	8.52	9.40	
Board D	Board D indicator (vs. A)	22.68	8.12	*
KS3mfine	KS3 maths fine grade	56.33	0.91	*
KS3efine	KS3 English fine grade	6.30	0.84	*
KS3sfine	KS3 science fine grade	19.08	1.06	*
Random variances and covariances		**Estimate**	**SE**	**Significant?**
Centre-level random variance		483.50	90.60	*
Candidate-level random variance		2,072.05	34.35	*

* = significant at 5% level

five or six times as large as for the OLS run, due to the fact that these estimates of board effects are seriously influenced by the clustering of candidates within centres, all of which take the same board. In this case the increased standard errors did not affect the significance of the results, but the use of OLS estimates could lead to incorrect conclusions about the standard errors of the between-board differences.

In Model 1 the centre-level variance in UMS scores (483.5) is 19% of the total variance (483.5 + 2072.1), showing that candidates at the same centre are more similar than candidates at different centres. Model 1 also assumes a fixed difference between examining boards, once KS3 attainment is taken into account. However, it may be reasonable to ask if the differences between boards vary for different levels of prior attainment. To answer this, we set up Model 2 in which we include interaction terms between examining boards and KS3 maths fine grade. To simplify the model, we include only maths fine grades – the other two core subjects do have an impact on UMS score, but this is small compared with KS3 maths.

Three interaction terms are included: **Bint, Cint** and **Dint**. In each case the term is equal to the board indicator (0 or 1) times the KS3 maths fine grade minus its mean (6.13). Thus a positive interaction term implies that the relation between outcome and KS3 fine grade is stronger for this board than the default (A), while a negative interaction implies the reverse. Results for Model 2 are shown in Table 4.

Table 4 Multilevel modelling results for Example 1 Model 2

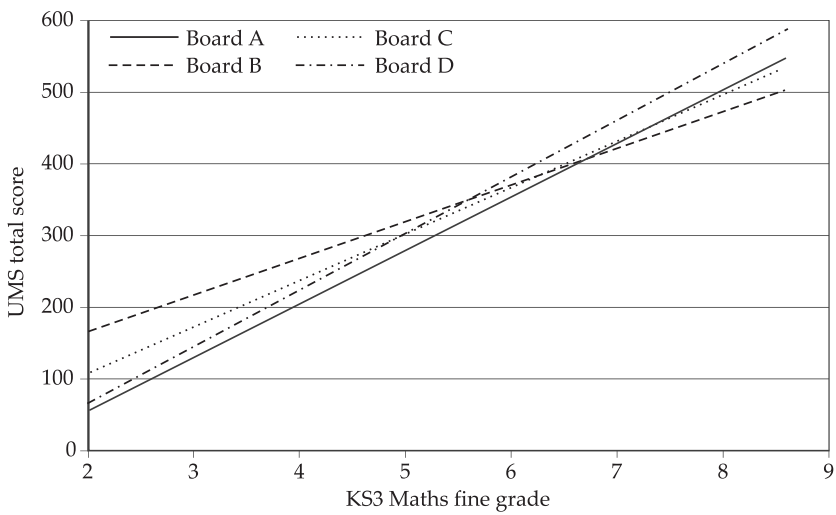
Variables		Estimates from modelling		
Name	**Description**	**Coefficient**	**SE**	**Significant?**
T3umstot	***UMS total score***	***Outcome variable***		
Cons	Constant term	–95.22	7.65	*
Board B	Board B indicator (vs. A)	14.12	7.15	*
Board C	Board C indicator (vs. A)	10.76	8.65	
Board D	Board D indicator (vs. A)	28.91	7.43	*
KS3mfine	KS3 maths fine grade	75.09	0.81	*
Bint	Board B × KS3mfine	–22.30	1.86	*
Cint	Board C × KS3mfine	–7.49	1.73	*
Dint	Board D × KS3mfine	5.70	1.36	*
Random variances and covariances		**Estimate**	**SE**	**Significant?**
Centre-level random variance		400.95	76.51	*
Candidate-level random variance		2,163.94	35.87	*

* = significant at 5% level

The main effects are similar to Model 1, with the Board C indicator non-significant (and Board B only borderline significant at the 5% level). However, all three interactions are statistically significant, including Board C. In two cases (Boards B and C) the interaction is negative; for Board D it is positive. The combined effects of the main effects and interactions from this model are illustrated in Figure 3, which shows a plot of the expected UMS scores for different values of KS3 fine grade for each board.

In this case the centre-level variance is 16% of the total variance, implying that part of the difference between centres can be explained by the examining board interactions.

Figure 3 Expected UMS scores for each board from Example 1 Model 2



This model could be developed in a number of different ways, but one possible addition is to consider the possibility that the relationship between UMS score and KS3 fine grade may vary from centre to centre, as it appears to vary from board to board. To model this possibility we need to make the coefficients of KS3 maths fine grade random at the centre level. If we do this, we get the fitted Model 3 results as shown in Table 5.

Table 5 Multilevel modelling results for Example 1 Model 3

Variables		**Estimates from modelling**		
Name	**Description**	**Coefficient**	**SE**	**Significant?**
T3umstot	***UMS total score***	***Outcome variable***		
Cons	Constant term	–93.57	17.49	*
Board B	Board B indicator (vs. A)	12.72	6.52	
Board C	Board C indicator (vs. A)	11.03	7.89	
Board D	Board D indicator (vs. A)	27.85	6.73	*
KS3mfine	KS3 maths fine grade	74.80	2.53	*
Bint	Board B × KS3mfine	–23.35	3.60	*
Cint	Board C × KS3mfine	–10.05	4.19	*
Dint	Board D × KS3mfine	4.73	3.46	
Variances and covariances		**Estimate**	**SE**	**Significant?**
Centre-level				
	Random variance (intercept)	3,487.14	816.97	*
	Random variance (KS3mfine)	72.84	18.37	*
	Covariance (intercept and KS3mfine)	–481.50	120.34	*
Candidate-level random variance		2,122.87	35.32	*

* = significant at 5% level

Results are similar to Model 2, except that neither Boards B nor C are overall significantly different from A, and although the overall Board D effect remains significant this is not true for the Board D interaction term. The variance at the centre level in the coefficient of KS3 fine grade is estimated as 72.8, equal to a standard deviation of 8.5, whereas the overall average coefficient is 74.8. This implies there is a reasonable amount of variation between centres in the relationship between prior attainment and GCSE results, and not taking this into account can change the conclusions of the comparability study.

So what have we learned from this example comparability study using separate forms? In terms of the differences between examining boards the following was found:

- There are overall differences in the results obtained for certain boards (controlling for KS3 prior attainment) and those for Board A; there is clear evidence of this for Board D and less clear evidence for Board B.

- The relationship between KS3 prior attainment and final outcome also varies between examining boards; Boards B and C have significantly less strong relationships than Board A.

- This relationship also varies between centres, and this variation if not taken into account may affect the conclusions of our comparability study.

These results should be regarded as indicative, based on the final model fitted. Other models, or the inclusion of more background information, may change these conclusions. In terms of what we have learned about the modelling process, we can say the following:

- OLS regression may give similar coefficients to those obtained from multilevel modelling, but is likely to underestimate the standard errors if within-centre clustering is not taken into account. This can affect the conclusions of comparability studies.

- Interactions to study differential effects for different boards relative to prior attainment are an important element of such studies and should be included in the model.

- Random coefficients at the centre level can be fitted in multilevel modelling, and these can be informative and affect the conclusions of the study.

Interaction terms are, of course, not restricted to multilevel modelling and can be fitted in other types of model, including OLS. When such models are used it is important to take account of what they mean in terms of 'comparability'. In essence we are saying that comparability needs to be assessed not just at a single point on the prior attainment scale, but at every point. Two boards may appear comparable on average, but if one produces higher scores for lower-attaining pupils than the other, and vice versa for higher-attaining pupils, then comparability is not achieved.

One option is to use the fitted model to standardise or adjust the results of different boards onto a consistent scale. This would be relatively straightforward when allowing for board effects, but it is not clear to what extent random coefficients for each centre should be allowed for. In the main, the results of such comparability studies are not used to adjust marks or grades retrospectively, but to inform the grade-setting process for the next round.

Other examples of multilevel modelling applied to this 'single forms' scenario can be found in Baird & Jones (1998) and Pinot de Moira (2000; 2002a).

4.2 Example 2: Multiple forms

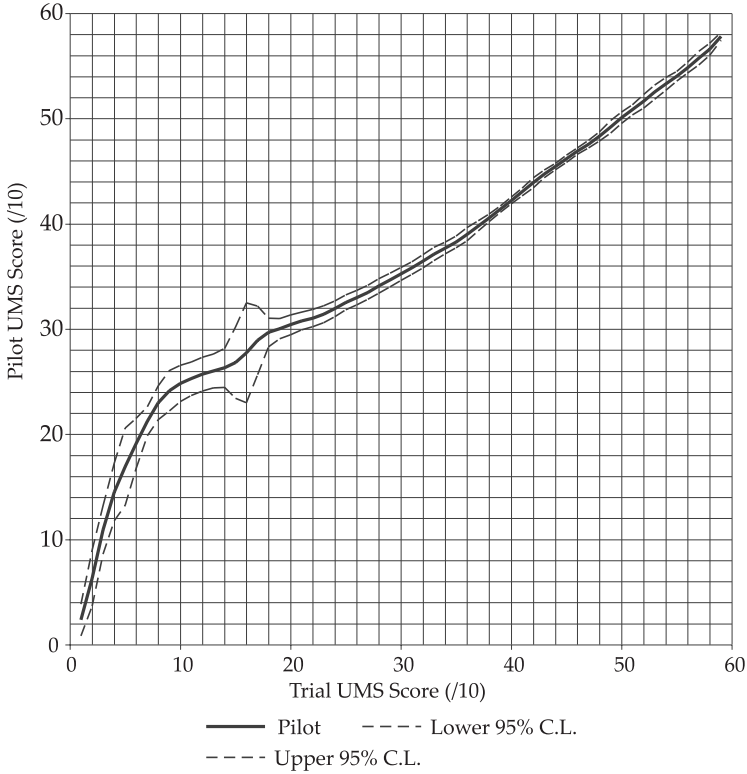
In the previous example no candidate took more than one examination, and the comparison between boards had to be based indirectly on the relationship with prior attainment. It would seem in principle more powerful to be able to compare boards, or forms of examination, directly by getting candidates to take more than one form so that unmeasured differences between candidates can be controlled for. As ever, there

are drawbacks with this approach: candidates may be differentially motivated between forms, or there may be an effect due to the order in which forms are taken or because of a time gap between taking them. In any case, we need to assume that the candidates who take both papers are a representative sample of the appropriate population. However, data from 'multiple forms' trials can be analysed in a powerful way using multilevel modelling, subject to these and other caveats.

As an example for this analysis we shall also take data from the study into alternate forms of the GCSE mathematics examination carried out in 2005 (see Stobart *et al.*, 2005), but in this case we will consider data from candidates who attempted two different alternatives to the three-tier paper, the so-called 'Pilot' and 'Trial' structures. In this case the candidates were all from the same board: 7,146 attempted the Pilot version, 732 the Trial version and 695 did both.

Here we have in some ways a more powerful data set for investigating differences in standards for two or more different forms of a test, because we have data on identical individuals who have attempted more than one form. Perhaps the most straightforward way of comparing standards is to carry out equipercentile equating using the 695 candidates who did both versions (see Stobart *et al.*, 2005) – Figure 4 shows the resulting equating graph. Note that to simplify the equating procedure we have divided the UMS score by ten in both forms.

Figure 4 Equating Pilot and Trial versions using common candidates



This is quite informative, but the estimated confidence intervals (CI) are based on a relatively simple formula and take no account of clustering within centres. We shall therefore explore this relationship further, taking account of all available data including background information on KS3 fine grades and the centres in which the examinations were taken.

In this case the kind of model we want is multivariate (see Goldstein, 2003, pp. 139ff), because candidates have more than one outcome to be modelled. We therefore introduce a lower level below the candidate for a version indicator (1 = Pilot, 2 = Trial) to enable this to be modelled. In addition we introduce separate indicators (0/1) for both Pilot and Trial, and include both in the model with separate random variances. No constant term is included in this case, as we have separate intercepts for the two forms and a constant would make the model over-determined. Table 6 shows the results for this Model 1 with no background factors.

Table 6 Multilevel modelling results for Example 2 Model 1

Variables		**Estimates from modelling**		
Name	**Description**	**Coefficient**	**SE**	**Significant?**
Umstot	***Total UMS score***	***Outcome variable***		
Pilot	Indicator for Pilot version	342.73	6.87	*
Trial	Indicator for Trial version	296.16	8.75	*
Random variances and covariances		**Estimate**	**SE**	**Significant?**
Centre-level				
	Pilot variance	2,488.11	500.60	*
	Trial variance	2,265.31	717.22	*
	Pilot/Trial covariance	2,304.91	554.32	*
Candidate-level				
	Pilot variance	10,007.64	170.84	*
	Trial variance	20,674.04	848.43	*
	Pilot/Trial covariance	11,772.58	369.73	*

* = significant at 5% level

From this it is clear that the two forms have different overall means (342.7 and 296.2) but similar between-centre variances (2,488 and 2,265). The within-centre variances are rather different, with the Trial having over twice the variance between candidates of the Pilot.

Relative relationships with prior attainment were modelled by including KS3 maths fine grade for both forms, plus an interaction term to see if the relationship was different for the Trial (interaction term = 0 for Pilot, and KS3 fine grade minus 5.9 for the Trial)[8]. However, Figure 4 indicates a possible non-linear relationship between Pilot and Trial, so non-linearities in the relationship of each with KS3 fine grade were also included in the model.

The full set of variables included in the model is therefore:

Pilot Indicator for Pilot version (random at centre level)
Trial Indicator for Trial version (random at centre level)
KS3mfine KS3 mathematics fine grade (fixed effect)
Verk3int Interaction between version and KS3 fine grade. Set to zero for the Pilot, and equal to fine grade value minus 5.9 for the Trial (fixed effect)
KS3msq Square of KS3 mathematics fine grade (fixed effect)
Verk3sq Interaction between version and fine grade squared. Set to zero for the Pilot, and equal to (fine grade minus 5.9) squared for the Trial (fixed effect)

Results for this model are shown in Table 7.

Table 7 Multilevel modelling results for Example 2 Model 2

Variables		Estimates from modelling		
Name	**Description**	**Coefficient**	**SE**	**Significant?**
Umstot	***UMS total score***	***Outcome variable***		
Pilot	Indicator for Pilot version	–181.89	12.81	*
Trial	Indicator for Trial version	–237.25	15.19	*
KS3mfine	KS3 maths fine grade	108.17	4.38	*
Verk3int	Version × KS3mfine	15.38	2.81	*
KS3msq	KS3mfine squared	–2.67	0.37	*
Verk3sq	Version × KS3msq	7.95	2.03	*
Random variances and covariances		**Estimate**	**SE**	**Significant?**
Centre-level				
	Pilot variance	321.72	68.19	*
	Trial variance	633.23	374.56	
	Pilot/Trial covariance	333.89	161.20	*
Candidate-level				
	Pilot variance	2,793.78	47.72	*
	Trial variance	10,435.20	519.60	*
	Pilot/Trial covariance	3,196.03	172.75	*

* = significant at 5% level

In this case the interaction terms are statistically significant, implying the relationship with prior attainment is different for the two forms. The between-centre variance for the Trial is now not significant, but the residual within-centre variance for the Trial is almost four times that for the Pilot. The non-linear terms in KS3 mathematics fine grade are significant, although the form of the non-linearity is different for the two versions. Figure 5 illustrates Model 2 in terms of the expected UMS score for each form as a function of KS3 fine grade.

From the above, there is clearly a mismatch between the two forms for much of the prior attainment range. This is consistent with Figure 4, although taking prior

attainment into account clarifies where the main mismatch is. Putting the relationship with KS3 fine grade random at the school level gives Model 3, shown in Table 8.

Figure 5 Expected UMS scores for each form from Example 2 Model 2

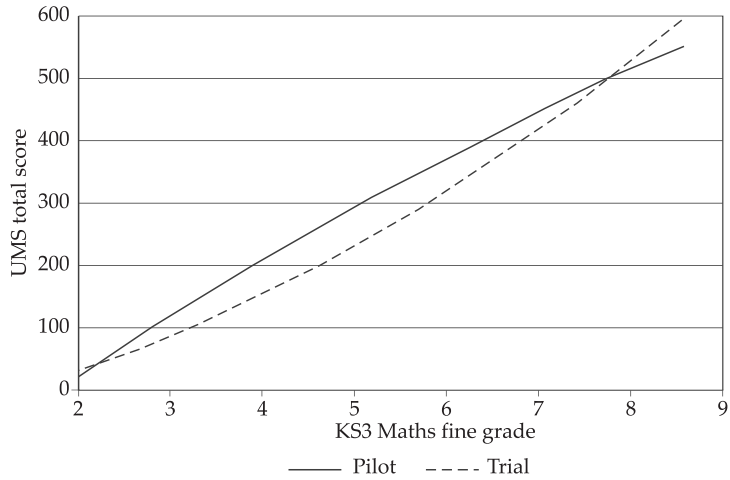


Table 8 Multilevel modelling results for Example 2 Model 3

Variables		Estimates from modelling		
Name	**Description**	**Coefficient**	**SE**	**Significant?**
Umstot	***Total UMS score***	***Outcome variable***		
Pilot	Indicator for Pilot version	–162.79	15.74	*
Trial	Indicator for Trial version	–212.49	17.69	*
KS3mfine	KS3 maths fine grade	103.19	4.86	*
Verk3int	Version × KS3mfine	16.16	2.81	*
KS3msq	KS3mfine squared	–2.41	0.41	*
Verk3sq	Version × KS3msq	8.00	2.03	*
Random variances and covariances		**Estimate**	**SE**	**Significant?**
Centre-level				
	Pilot variance	3,440.18	823.15	*
	Trial variance	4,530.73	1,377.34	*
	KS3mfine variance	72.88	18.71	*
	Pilot/Trial covariance	3,848.62	1,003.78	*
	Pilot/KS3mfine covariance	–487.56	122.60	*
	Trial/KS3mfine covariance	–562.84	150.95	*
Candidate-level				
	Pilot variance	2,733.72	46.85	*
	Trial variance	10,407.86	519.01	*
	Pilot/Trial covariance	3,149.62	171.17	*

* = significant at 5% level

So what have we learned from this example comparability study using multiple forms?

- There are significant differences between results obtained in the Pilot and Trial forms, in terms of the relationship with prior attainment. For example, a candidate with a KS3 fine grade of 6.0 would be expected to score 371 on the Pilot and 317 on the Trial.
- There are significant non-linearities in the relationship with prior attainment, and these vary between forms.
- Only at the highest and lowest levels of prior attainment are the two forms approximately comparable.
- Variation between candidates, controlling for prior attainment, is much higher for the Trial form than for the Pilot.
- The correlation between Pilot and Trial versions at the centre level is 0.975, implying that centres tend to perform overall equally well or equally poorly on both versions.
- The correlation between Pilot and Trial versions at the candidate level is 0.590, implying that there is a less strong relationship between the two versions for individual candidates.
- As a result of these differences we would be forced to conclude that the two forms were not directly comparable.

In terms of what we have learned about the modelling process, we may say:

- Multivariate models require an extra, lowest level to allow multiple outcomes per candidate.
- Separate indicators for each form can be used, random at all levels above the lowest, and the constant terms should then be omitted.
- Graphs of expected outcome as a function of prior attainment for the different forms may be a powerful way of illustrating the comparability or otherwise of different forms.

In a later section we will return to this data set when we consider complex variance models.

5 Modelling different outcomes

In the previous section we examined different ways of modelling the structure of the data in comparability studies, but throughout we assumed that the outcome of interest could be treated as a continuous numerical variable. However, this is often not legitimate – examination results can be reported as ordered categorical outcomes (grades or levels) or just as a binary outcome (pass or fail). The linear models with

Normally distributed error terms that we have used so far in this chapter are inappropriate for such outcomes, and in this section we will describe suitable models, using initially an example from previous literature in the area.

Probably the earliest attempt to compare examination 'standards' using multilevel modelling and the 'catch all' definition was the work of Baird & Jones (1998). Other more recent studies working in a comparable way include Pinot de Moira (2000, 2002a).

In their paper Baird & Jones (1998) compare three different statistical techniques in the analysis of an inter-board comparability study on 1996 GCSE art and design (Unendorsed) grades, which was undertaken by the boards themselves on behalf of the Joint Forum for the GCSE and GCE (Jones *et al.*, 1997). They concluded that ordered logistic multilevel modelling was the best option, but that it still failed to deal with the fundamental problems. To quote from their report:

> It is argued that ordered logistic multilevel modelling is the most appropriate of the three forms of statistical analysis for comparability studies using examination grade as the outcome variable. Although ordered logistic multilevel modelling is considered an important methodological advance on previous statistical comparability methods, it will not overcome fundamental problems in any statistical analysis of examination standards. It is argued that ultimately examination standards cannot be measured statistically because they are inextricably bound up with the characteristics of the examinations themselves and the characteristics of the students who sit the examinations.
>
> Baird & Jones (1998)

The Baird & Jones (1998) study is now described in some detail to make the method clear and to highlight features of interest. A random sample of approximately 1,500 art and design candidates from each of the four English GCSE examining boards, stratified by centre type, was sent a questionnaire, designed to measure a few of the key variables expected to have a significant relationship with awarded grades. There was an approximately 33% response rate to this questionnaire. The variables used in the analyses included individual responses and responses aggregated to examination centre level. Variables found to have statistically significant effects in the analyses were measures of pupil attitudes, plans, gender and background. It was not possible to obtain a measure of prior attainment for the individual pupils, but an aggregated ability measure from school league-table information was included.

Three different kinds of statistical methodology were used in the analysis of this project. These were:

1. Ordinary Least Squares (OLS) linear regression at candidate level treating the grade outcome as a continuous variable.

2. Linear multilevel modelling treating the grade outcome as a continuous variable.

3. Ordered logistic multilevel modelling considering whether the candidates succeeded at various grades within the examination. Thus a candidate who gains

a B grade will be considered as having also gained a C, D or E grade, but not an A grade. See below for a fuller treatment.

The first two analyses are largely similar to those described previously, so we will not describe these models in detail. The main focus of interest for us is on the third model, the ordered logistic multilevel model.

In an ordered logistic model, examination grades are treated as ordered categories, instead of as numerical values. In this type of analysis, an equation is first found for the probability that a case is above the first (lowest) category. Following this, an equation is found for the probability that the case is above the second lowest category and so on. The response variable in the ordered logistic regression is the cumulative grade (s) for each candidate *i* in teaching group *j* in centre *k*,

$$\log it(\pi_{ijk}^{(s)}) = \beta_0^{(s)} + \sum_{p=1}^{P} \beta_p x_{pijk} + \sum_{q=1}^{Q} \beta_q z_{qijk} + v_k^{(s)} + u_{jk}^{(s)} \qquad (13)$$

where:

$\pi_{ijk}^{(s)}$ = the probability the examination grade is *s* or better

$\beta_0^{(s)}$ = the intercept for the particular grade *s*

β_p = the coefficient for the p^{th} explanatory variable

x_{pijk} = the value of the p^{th} explanatory variable for candidate *i* in teaching group *j* in centre *k* for the particular grade *s*

β_q = the effect of board *q*

z_{qijk} = an indicator that candidate *i* in teaching group *j* in centre *k* is taking the particular board *q*

$v_k^{(s)}$ = the effect of centre *k* for the particular grade *s*

$u_{jk}^{(s)}$ = the effect of teaching group *j* in centre *k* for the particular grade *s*.

This model investigates the effect of the *qth* board, assuming this is uniform at all of the categories. If it is suspected that the difference between boards is greater at some categories than others then different values $\beta_q^{(s)}$ can be fitted. For a fuller discussion of multilevel models for discrete outcomes, see Goldstein (2003, pp. 95ff).

Table 9 shows the results from the three types of analyses side by side. Only an extract of the results from these tables is shown here: other aspects of the analysis, not referring directly to the inter-board comparability, are not shown here. Interested readers should consult Baird & Jones (1998). Four boards are compared. As before, in

each case, Board 0 is taken as a reference category, and the results for the other boards are expressed in comparison with these.

The first two columns relate to an OLS analysis. The first column shows the value for the difference (Board *q* vs. Board 0), and the second column shows the standard error for this. An impression of the probability value for these comparisons can be gained by dividing the difference by the corresponding standard error, and comparing the result with the 0.05 level for a two-tailed z-distribution. It can be seen that Board 3 appears to have a lower value than Board 0, while there is no statistically significant difference between Board 0 and the other two boards (Boards 1 and 2). The figures here relate to the contrasts between Board 0 and the other three boards. There are of course other contrasts that could be considered, such as Board 1 vs. Board 2. While this has not been considered specifically it seems quite likely that the contrast between Board 3 and Board 1 would also be statistically significant. This could be easily investigated, if required. Comparisons using this method are likely to give biased estimates of the statistical significance however, since candidates taking their exams within a single centre are likely to be more similar to each other than are candidates chosen completely at random.

This problem can be met by the use of multilevel modelling techniques. The results of this are shown in the third and fourth columns of figures. The standard errors are all substantially larger than those estimated for the OLS analysis, and largely as a result, while Board 0 is the lowest, none of the inter-board differences are statistically significant. In fact, the fitted constants are also different from the results in the OLS analysis.

Table 9 Comparison of different models on inter-board comparability results

Board (vs. Board 0)	OLS results		Multilevel modelling (linear) results		Multilevel modelling ordinal results for Grade A*		Multilevel modelling ordinal results for Grade A	
	Coeff	SE	Coeff	SE	Coeff	SE	Coeff	SE
Board 1	0.03	0.09	–0.16	0.20	–1.87	0.59	–0.81	0.34
Board 2	–0.02	0.08	–0.06	0.19	–0.94	0.38	–0.90	0.30
Board 3	–0.17	0.08	–0.16	0.20	–1.17	0.40	–0.63	0.31

Such results are of general eye-catching interest, but may not be the best way to look at possible differences. Examination results are awarded by grade and it is generally considered that the grade awarded is accurate to plus or minus one grade (Newton, 2005). Consequently, the main focus should be on grade borderlines. If there is a difference between the standards set at just one borderline it is conceivable that this could give a statistically significant overall mean difference. The process of awarding a grade combines two processes: first awarding a mark to a script, and then comparing the mark awarded with the grade boundaries. Only key grades are considered at award meetings – the rest are set arithmetically. If one examining board

finds that its grading is out of line with others, then it will be concerned to find out whether this is due to boundary decision-making between grades.

The next analysis reported therefore treats the grade outcome as an ordered categorical variable. Only the results for the A* borderline and the A borderline are shown in this chapter, since the inter-board differences at other grade boundaries were not statistically significant. The last four columns in Table 9 show the results for these. Taking one example for illustration purposes, the fitted constant for Board 1 for grade A* is negative, and more than twice its standard error. This figure relates to the log-odds, ceteribus paribus, of getting this grade from Board 1, compared with those of getting this grade from Board 0. Transforming to a more intuitive metric, this means that the odds of getting a result this good in Board 1 are only 15% of those in Board 0. However, it should be noted that these results are subject to large margins of error, and that analysis of a data set with good measures of prior attainment at the individual candidate level might provide different results.

In this section so far we have reported results based in practice on logistic modelling, where the outcome of interest is a single binary variable (i.e. does the candidate get grade A* or above, or not?). When looking at comparability over several grade boundaries, this approach requires the application of a separate model for each grade. Another approach is to use an ordered categorical multinomial model (see Goldstein, 2003, pp. 101ff). An example is taken from Stobart *et al.* (2005), and considers whether candidates with different levels of KS3 attainment have different probabilities of getting higher grades in the Pilot compared with the Trial version of the examination, or vice versa.

An ordered categorical multinomial model (see Goldstein, 2003, p. 104) was fitted (see also equation (13)), looking at three categories:

1. Trial grade higher than Pilot grade
2. grades the same on two tests
3. Pilot grade higher than Trial grade.

The default category was taken as the first ('Trial>pilot'), and a model with constant parameters for KS3 fine grade was fitted, with two levels – school and candidate. Essentially we are fitting two linked logistic models: one for the grades being the same on the two tests ('Same'), and the other for the Pilot grade being higher than the Trial ('Pilot>trial'), with the same relationship assumed with prior attainment in both cases. The full fitted model is shown in Figure 6. This is an example of a more complex equation window from MLwiN, and one whose features may need more time to understand. Model parameters are also displayed in tabular form in Table 10.

Figure 6 Ordered multinomial multilevel model fitted to Trial and Pilot data

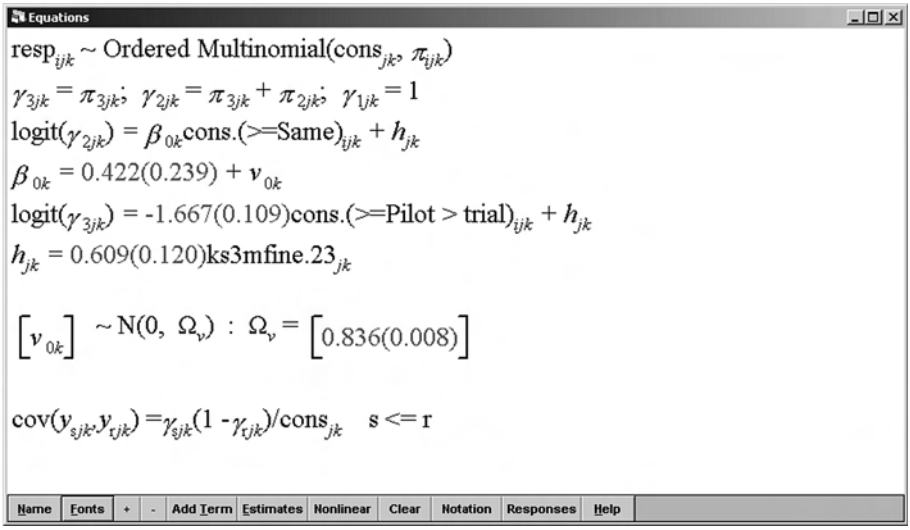


Table 10 Multilevel modelling results for ordered multinomial multilevel model fitted to Trial and Pilot data

Variables		Estimates from modelling		
Name	**Description**	**Coefficient**	**SE**	**Significant?**
Resp	*Three-category response*	*Outcome variable*		
Cons. (>=Same)	Constant term for contrast between categories 2 and 1	0.4217	0.2388	
Cons. (>=Pilot>trial)	Constant term for contrast between categories 3 and 2	–1.667	0.1092	*
KS3mfine	KS3 maths fine grade (centred on 6.0)	0.6094	0.1196	*
Random variances and covariances		**Estimate**	**SE**	**Significant?**
Centre-level				
	Random variance	0.8357	0.0085	*

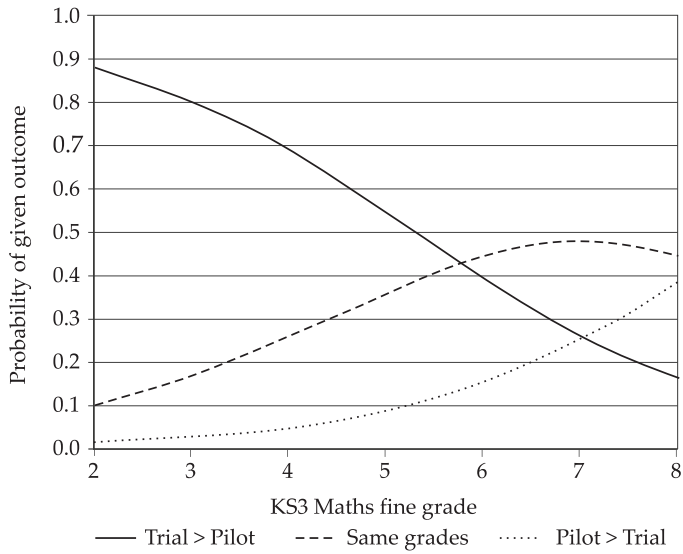
Two categories are explicitly featured in the model: 'Pilot>trial' and 'Same'. The modelled probability for the latter includes the former. In both cases there is a relationship with prior attainment ('ks3mfine') which has the same slope (0.609) in the logit metric.[9]

The log-odds for 'Pilot>trial' has a constant estimate of –1.667, whereas for the log-odds of 'Same' or 'Pilot>trial' the constant estimate is 0.422. The coefficient of *ks3mfine* (centred on the value 6.0) is 0.609. Substituting a value of 0.0 for *ks3mfine*, we estimate the two log-odds as –1.667 and 0.422 respectively, with corresponding

probabilities of 0.159 and 0.445. By subtraction, we find: P['Pilot>trial'] = 0.159; P['Same'] = 0.445; and P['Trial>pilot'] = 0.396.

In this case there is a clearly significant random variance at the school level, implying that the relationship between Pilot and Trial grades does vary from school to school. Expected results of this model, controlling for KS3 results, are shown in Figure 7.

Figure 7 Probabilities for Pilot and Trial comparison from ordered multinomial model



This model illustrates that for candidates with lower prior attainment, the Trial structure seems to be advantageous as they have a higher chance of getting a better grade on this than on the Pilot. As prior attainment increases, the two systems become more balanced and the apparent advantage of the Trial disappears.

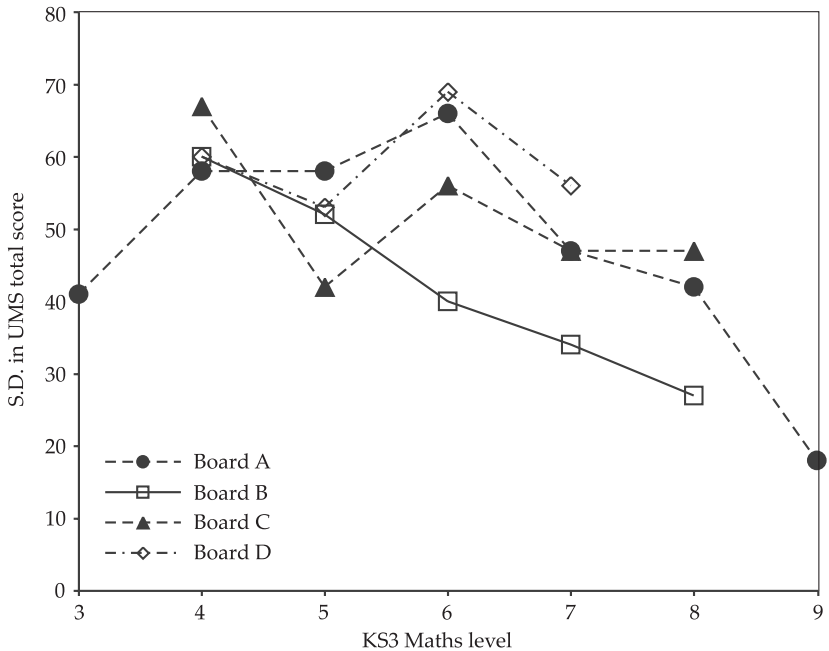
This kind of ordered categorical model is very powerful, and in principle should be used more widely. The truth is that our most common examination outcomes (grades or levels) are actually ordinal in nature, although much of the quantitative analysis carried out tends to treat them as if they were numerical interval scales.

6 Complex variance models

In the previous two sections we have dealt in some detail with two important aspects of model selection in comparability studies – modelling the structure of the data and the outcome of interest. It may seem that modelling the variance structure of the data is of less importance, but in certain cases false assumptions about this can lead to incorrect conclusions from the study.

Let us take as an example the data for Model 1 from section 4, with UMS scores from four examining boards. In Figure 8 we plot the standard deviation in the UMS score as a function of the KS3 mathematics level for each board. From this there is some evidence that the standard deviation (and therefore the variance) in UMS score is not constant across the KS3 prior ability range, as has been assumed in all the models to date. It is not clear whether this will affect the conclusions of the modelling, but in a comparability study this should be checked by including this feature of the data in the model.

Figure 8 Standard deviation in UMS score versus KS3 level



Goldstein (2003, pp. 63ff) shows how this may be done, by suitable modification of our models. By introducing a random coefficient of the relevant background variable at the lowest level of the model, we can generate a quadratic function of this variable as a model for the pupil-level variance.[10] Note that although some of the coefficients of this function may be described as variances in the output from the software, they are not and need not be constrained to be non-negative – the true pupil-level variance is defined by the whole function, not by any of its elements individually. This means that non-negativity constraints on variance parameter estimation in the software used need to be relaxed when complex variances are being fitted.

We will go back to our examples from section 4 to show how these complex variances can be fitted, starting with Model 1. Table 11 shows the results of setting prior attainment (*ks3mfine*) random at the candidate level.

Table 11 Complex variance model fitted to Example 1 data

Variables		Estimates from modelling		
Name	**Description**	**Coefficient**	**SE**	**Significant?**
T3umstot	***Total UMS score***	***Outcome variable***		
Cons	Constant term	–107.63	17.40	*
Board B	Board B indicator (vs. A)	12.78	6.43	*
Board C	Board C indicator (vs. A)	10.79	7.77	
Board D	Board D indicator (vs. A)	28.18	6.61	*
KS3mfine	KS3 maths fine grade	76.99	2.58	*
Bint	Board B × KS3mfine	–36.08	3.71	*
Cint	Board C × KS3mfine	–11.37	4.23	*
Dint	Board D × KS3mfine	3.01	3.55	
Random variances and covariances		**Estimate**	**SE**	**Significant?**
Centre-level				
	Random variance (intercept)	3,555.67	838.27	*
	Random variance (KS3mfine)	78.83	19.53	*
	Covariance (intercept and KS3mfine)	–506.46	125.75	*
Candidate-level				
	Pseudo variance (intercept)	–253.30	690.67	
	Pseudo variance (KS3mfine)	–135.66	15.90	*
	Pseudo covariance	625.08	106.00	*

* = significant at 5% level

Comparison with Table 5 shows little difference in terms of the main coefficients and the substantive findings of the modelling. The candidate-level variance matrix now shows apparently negative variances, but as mentioned above this is not problematic, as these are not real variances but coefficients in the variance equation:

$$\textit{Pupil-level variance} = -253.30 + 2*625.60*ks3mfine - 135.66*ks3mfine^2 \quad (14)$$

Figure 9 illustrates the model standard deviations as a function of KS3 fine grade.

A similar complex variance model was fitted to the Example 2 data, by modifying Model 3 (Table 8) to allow *ks3mfine* to be random at the candidate level – results are shown in Table 12.

Again, there is little change in the main coefficients, although this time the diagonal terms of the candidate-level variance matrix are all positive. The effects in terms of standard deviation of outcomes as a function of prior attainment are shown in Figure 10, where the variance models are different for the Pilot and Trial outcomes.

Figure 9 Model candidate-level standard deviation as a function of KS3 level

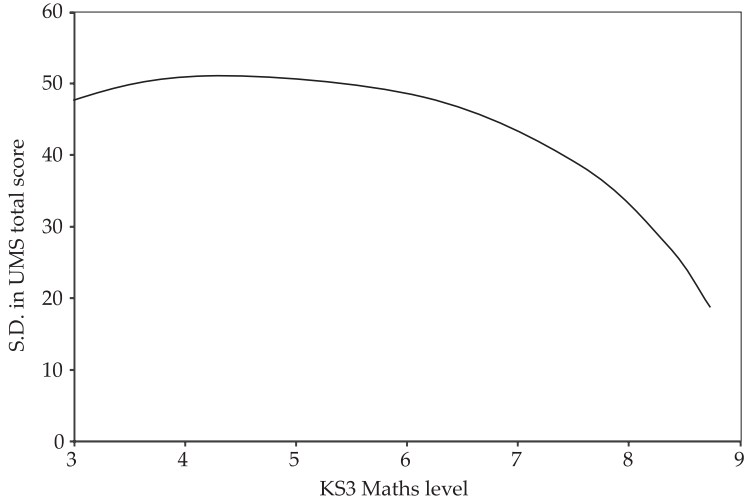
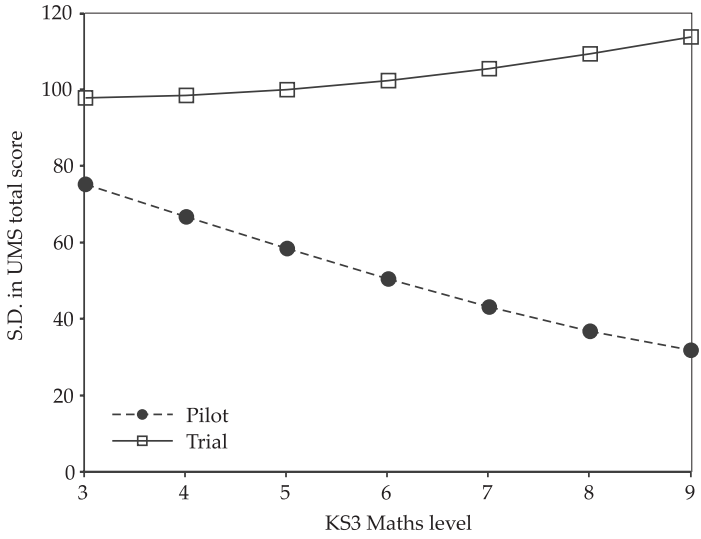


Table 12 Complex variance model fitted to Example 2 data

Variables		Estimates from modelling		
Name	**Description**	**Coefficient**	**SE**	**Significant?**
Umstot	***Total UMS score***	***Outcome variable***		
Pilot	Indicator for Pilot version	–174.66	16.79	*
Trial	Indicator for Trial version	–224.98	18.73	*
KS3mfine	KS3 maths fine grade	107.60	5.05	*
Verk3int	Version × KS3mfine	15.71	2.77	*
KS3msq	KS3mfine squared	–2.82	0.41	*
Verk3sq	Version × KS3msq	8.86	1.93	*
Random variances and covariances		**Estimate**	**SE**	**Significant?**
Centre-level				
	Pilot variance	3,561.67	868.31	*
	Trial variance	4,696.08	1,473.47	*
	KS3mfine variance	79.00	20.17	*
	Pilot/trial covariance	3,976.98	1,061.92	*
	Pilot/KS3mfine covariance	–516.00	130.78	*
	Trial/KS3mfine covariance	–592.32	161.35	*
Candidate-level				
	Pilot (pseudo) variance	10,299.57	1,015.93	*
	Trial (pseudo) variance	10,208.55	1,601.44	*
	KS3mfine (pseudo) covariance	86.87	24.47	*
	Pilot/trial (pseudo) covariance	6,758.68	1,136.35	*
	Pilot/KS3mfine (pseudo) covariance	–907.79	159.03	*
	Trial/Ks3mfine (pseudo) covariance	–240.69	170.78	*

* = significant at 5% level

Figure 10 Model 2 standard deviations as function of KS3 level (modelled)



The assumption of constant variance is one which is frequently made when modelling comparability study data, but it should always be critically evaluated, either by exploratory data analysis and/or by allowing for non-constant variance in the modelling. If this is not done, there is the potential for misleading results to be obtained and for the findings of the model study to be challenged.

7 Review of existing studies

In this section we will pause in our development of multilevel techniques in order to summarise and review existing comparability studies that have used this approach in order to address Cresswell's (1996) 'catch-all' definition, described in the introduction. This is a pretty ambitious definition, but there have been a number of studies that have at least tried to model its requirements statistically, though Baird & Jones (1998) admit that 'research can only approximate the "catch-all" definition, as the researcher does not have access to measurements of all of the factors which influence examination performance'. Problems and issues are discussed in detail in Baird & Jones (1998) so they will be touched on only briefly here, using their terminology. First, there is the problem of measurement, one aspect of which is the assumption that the relation between the ability measure and the examination outcome is the same for each syllabus or board. Second, there is the problem of interactions of independent variables with syllabus or board, when, for example, girls do better in one examination than boys, but not in another. The third problem touched on is that of extrication: it could arise if one syllabus or board were more attractive than another, and increased the motivation of the pupils involved, so that difficulty and motivation were confounded.

A number of studies have been conducted using multilevel modelling by or on behalf of the examining boards or the QCA. They are too numerous to mention individually in the relatively short space of this chapter, but a selection showing the techniques involved is now considered. The studies we cover here include Baird & Jones (1998), Pinot de Moira (2000; 2002a; 2003), Jones *et al.* (1997), Tymms & Vincent (1995), Bell & Dexter (2000). We compare how effective these are likely to be in operationalising the Cresswell definition. Two main aspects are considered: first the techniques employed, and second the predictor variables used.

7.1 Techniques used

In reviewing this type of work we consider that it is especially important to bear in mind the practical applications as well as any theoretical considerations, and a comment on how boards arrive at final grades is relevant. Each examination for each board is scored to give a total number of marks (although there are some recent developments in the use of IRT methods). Each board then organises meetings to convert these marks into grades, which are the major currency in the public examination world. These meetings determine the boundary point in terms of score for major cut-points: for example, F/G, C/D and A/B for GCSE. An interpolation procedure is then used in determining the grades between these boundaries.

Baird & Jones (1998), as described earlier (Table 9), compared three techniques in their investigation: Ordinary Least Squares at candidate level, treating the (ordinal) grade outcome as a continuous variable; a multilevel model also treating grades as continuous; and a multilevel ordinal multinomial model. In the two multilevel analyses, candidates were nested within teacher groups, which were nested within centres. Baird & Jones did not investigate the possibility of doing a series of dichotomous outcome logistic investigations, stating only that 'Ordered logistic modelling is the most appropriate method for the analysis of examination grades' (p. 15). Bell & Dexter (2000), however, also compared a series of binary ordered logistic models.

Probably the main objection to multilevel modelling (MLM) on the part of critics is that it is more complex to execute and understand than OLS. It is generally considered that where OLS and MLM results differ, the latter are to be preferred. In their example, Baird & Jones showed that the continuous multilevel model was preferable to the OLS model since OLS found a difference in level between boards, but no difference in spread, while linear MLM found the opposite – no difference in levels, but differences in spread. Since, as noted earlier, the main focus of examination results, both in reporting and in 'fixing' any apparent problems lies in the actual grades awarded, it makes good sense to concentrate on grades and use an ordinal model. This model focused the locus of the differences to one particular boundary, that between A and A* for one board. This seems to

show that ordinal regression is superior to continuous variable regression in this instance.

Another possibility is that, instead of carrying out a single ordinal variable regression including all of the levels, one carries out a series of binary logistic analyses, concentrating on each of the relevant boundaries in turn. Bell & Dexter (2000) considered that this approach was preferable because 'the results are much simpler and more interpretable for less experienced users'. In general, a principle in statistics is to borrow strength from adjacent observations to supplement sparser data. It should be noted that that using a single ordered model rather than a set of binary splits means that we have a single random effect for each centre rather than a series of unrelated ones from each binary analysis that may not even be mutually consistent. However, in this example, where a substantial amount of data is available and we are not certain that mechanisms will be the same from one boundary to another, we agree that it can make more sense to carry out the analyses as parallel but separate exercises.

Comparative analyses assume that the 'same' process is taking place for all examination entrants. In practice, however, some cases appear to be outliers from the rest. In some instances the data are obviously errors, for example mis-transcriptions, but in other situations they may arise because they represent a different population or process. Errors and missing data may be corrected if it is obvious what the correct value should be. Otherwise the missing or erroneous data may be imputed, or the whole case excluded. Either of these procedures has to be carried out with care. Alternatively, it may be that such apparently anomalous outliers do not represent a problem with the recording of the data, but are indicative of the existence of a separate population. Bell *et al.* (2004) quote an example where apparently anomalous data arose as a result of 'mature' entrants. They found that it was possible to extend the model to take account of these, but the resulting model proved to be relatively complex. They found that simpler models arose by analysing the two populations separately.

7.2 Predictor variables included

This section has benefited from the theoretical papers of Bell & Dexter (2000) and Pinot de Moira (2002b). If the average performance on an examination for one board is higher than that on another, it is not necessarily the case that the first examination is 'easier'. The first possible explanation is that one group of pupils is simply better on that topic on the day, and that tested on another occasion, or in different circumstances, this difference might disappear.

Another alternative explanation is that the two groups of pupils taking the exams are different in some important relevant way. One can take account of this by carrying out some kind of regression, as described in this chapter, and including some measure of ability. The boards have carried out or commissioned a number of studies and we now describe the type of data used in a representative selection of these.

Bell & Dexter (2000) distinguish between prior, concurrent and subsequent attainment (though somewhat confusingly they refer to them all as 'outcomes'). They state that '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'. According to Bell & Dexter (2000), prior attainment could include Key Stage 3 scores for GCSE, or GCSE score for A level, while a concurrent measure could be a test to measure 'general ability', 'aptitude', or 'calibre', or a subject-based reference test, or a common element included as part of all examinations. Bell & Dexter also include subsequent attainment, but this may be thought to measure the rather different characteristics of usefulness and relevance.

Most studies aiming to use a measure of ability have used a measure of prior ability, though Dexter & Massey (2000) and While & Fowles (2000) have used some kind of concurrent measure of attainment. Each of these studies raised some questions[11]. While & Fowles used a common element of a mathematics tests. This was, however, a coursework aspect for all but one of the boards involved. This is unfortunate, in that it seems likely that coursework may be a less 'pure' measure of attainment, since there is the possibility of parental input to any such work. Further, even if it is accepted that this is not a problem, or at least that it is a comparable problem for all, the fact that one of the boards did not treat it as coursework means that this common element is not comparable for this board. Finally, and to an outsider really rather surprisingly, given that this was planned to be a comparability exercise, while four of the boards agreed a common grading, the remainder were unable to concur. Dexter & Massey used a calibration test containing verbal, spatial and numerical reasoning in a study comparing GCSE and IGCSE (International GCSE) results for a number of subjects. Newbould & Massey (1979, cited in Dexter & Massey, 2000) discuss whether to use a general or a subject-based calibration test, and advocate the former since the latter is more likely to be differentially biased against some of the syllabuses being compared.

The value of using a measure of attainment as a term in the regression is dependent on the degree of relationship between that measure and the outcome. A test of academic attainment or 'general ability' may well do a good job in allowing for prior differences for exams in (for example) English or mathematics, but is likely to be of less value for art or physical education. In the extreme, if there is no relation between the test and the outcome, the attempted adjustment will be ineffectual. Dexter & Massey suggest that this may be less of a problem than might first be imagined, since correlations between their general ability test and outcomes in the six 'most popular' examination subjects range between 0.56 and 0.74: even art and design grades correlated 0.53 with the calibration test, though it should be noted that this means that less than 30% of the variance in the examination grade is accounted for by the calibration test.

A further complication potentially arises if there is an interaction between examination type and ability. In this situation the difference between examinations will vary over the ability scale. A single 'difference' factor may be produced by producing an average of the differences, but it may be preferable to show the entire picture in graphical form (see, for example, Figure 3 in this chapter).

Additional complications arise if one tries to allow for other possible background factors. For example, if girls do better at GCSE on one topic than boys after allowing for KS3 attainment, and there are more girls on one examination than on another, what exactly are we doing if we allow for this statistically? It may be appropriate to reweight the sample so that equal proportions of boys and girls are present in the weighted sample, or include a gender term in the model, which has a comparable effect. There is a danger that doing so will disguise an important phenomenon, namely that boys who have done as well as girls up to Key Stage 3, are now finding themselves less successful. Is the examination biased against them, or is the curriculum in schools failing to hold their attention, or is there some other reason? Similar considerations apply also to such factors as ethnic background or type of centre attended (independent school, type of state school, etc.). It is important to bear in mind that simply because these techniques are statistical, this does not mean that they are all value-neutral.

Table 13 shows in tabular form the extent to which a range of 'typical' studies were able to meet the conditions of the Cresswell definition in terms of allowing for relevant background characteristics.

Most of the studies have some kind of individual-level measure of ability or attainment, though not all of these even have this: for example Jones *et al.* (1997) and Pinot de Moira (2000) have some correlates of attainment, such as reported age planning to leave school, and Baird & Jones (1998) have prior attainment measures aggregated to centre level only. It is very likely that the predictive ability of the different measures of prior attainment used in these studies will vary, especially between subjects (see Pinot de Moira, 2002a; Dexter & Massey, 2000). None of the studies appears to have any information on school entry policy (except for While & Fowles, who considered tiered entry), or the competence of their teachers.

A few of the studies make an effort to obtain some measure of motivation, though this can prove problematic. While & Fowles (2000) described some of the difficulties in attempting to produce a good measure of attitudes. Such information did not already exist, and they were forced to ask the examination centres to distribute questionnaires to candidates taking exams. In their own words, 'a high proportion of centres failed to distribute (the questionnaire), and some of the candidates who did return it gave some questionable responses, thus calling into question the reliability of the questionnaire data'.

Table 13 Comparability studies and the Cresswell definition

Subject	Level and date	Outcome scale	Ability measure	Motiv-ation	Other pupil variables	Refer-ences*	Notes
English	GCSE 1998	Dichotomous	Some correlates	Some	Some	(1)	
English	GCSE 1998	Dichotomous	KS3 English		Some	(1)	
English	GCSE 1998	Grades as continuous	KS3 total			(2)	
Business	A/Voc A 2002	Dichotomous	Mean GCSE		Male/female	(3), (4)	Also 2001
Chemistry	A level 2002	Dichotomous	Mean GCSE		Male/female	(3), (4)	Also 2001
Geography	A level 2002	Dichotomous	Mean GCSE		Male/female	(3), (4)	Also 2001
Health Care	AVCE 2002	Dichotomous	Mean GCSE		Male/female	(3), (4)	Also 2001
12 subjects	A level 1993	Grades as continuous	Mean GCSE 'ability'		Male/female	(5)	12 studies
Art	GCSE 1996	Grades as continuous	Some correlates	Yes	Male/female, FSM, homework	(6)	
Maths	GCSE 1998	Dichotomous	Common test element	Yes	Male/female, homework	(7)	Considered tiering; problems with response rates
French	GCSE 2004	Dichotomous	KS3		Male/ female	(8)	Convergence problems
9 subjects	GCSE	Grades as continuous	Anchor test		Male/female, language	(9)	
Art & Design	GCSE 1998	Continuous and ordinal	Aggregated measures	Some	Male/ female, FSM, homework	(10)	

Key to pupil variable abbreviations:

Male/female Male/female differences
FSM Eligibility for free school meals

***Key to references:**

(1) Pinot de Moira (2000)
(2) Bell and Dexter (2000)
(3) Pinot de Moira (2003)
(4) Pinot de Moira (2002a)
(5) Tymms and Vincent (1995)
(6) Jones, Baird and Arlett (1997)
(7) While and Fowles (2000)
(8) Al-Bayatti (2005)
(9) Dexter and Massey (2000)
(10) Baird and Jones (1998)

It is clear from Table 13 that none of these studies comes near meeting the criteria for properly assessing comparability under the Cresswell definition. Certainly

sophisticated models are being employed to carry out the analyses, but this is going to be of little practical value if the underlying data is weak. If this area of application is to continue, then the work so far should be used as a jumping-off point for devising more rigorous investigations. It may be that some aspects of the factors to be allowed for are less critical, but it will be important for credibility to provide a valid justification for not including these.

All the above studies were selected because of their use of multilevel techniques, but it can be seen that this approach to modelling is not a sufficient condition for a fully effective comparability study. It could be argued that it is a necessary condition, but the study as a whole needs to be carefully designed in order to meet the challenge set by the Cresswell definition of comparability.

8 Practicalities

In this section we shall deal with:

1. software packages
2. acquiring suitable data
3. pitfalls and problems.

8.1 Software packages

When running multilevel modelling it will normally be necessary to acquire a specialist software package that is capable of dealing with all the necessary complexities of the modelling. Some general-purpose packages (e.g. SPSS, SAS, S+ and STATA) are now starting to include multilevel modelling modules, and new software is continually appearing in this area. For example, the WinBUGS software (Bayesian inference Using Gibbs Sampling) provides flexible software for the Bayesian analysis of complex statistical models using Markov chain Monte Carlo (MCMC) methods[12]. There are two specialist packages that have been widely used over a long period: MLwiN and HLM.

MLwiN was developed by the Centre for Multilevel Modelling at the Institute of Education, University of London[13] (see Rasbash *et al.*, 2002; 2004). It provides data-entry and manipulation facilities, a graphical interface, a range of options for estimation, a facility for displaying models in terms of equations and fitted parameters (see screen shots within this chapter) and a command interface that replicates the structure of the earlier DOS-based version, MLn. The latest version, MLwiN 2.0, has a number of additional features including the ability to fit more complex models, such as ordered and unordered multinomial models. The program can fit up to five levels of hierarchy, and the size of data set that can be handled appears to be limited only by the capacity of the machine on which it is run. The authors have run four-level models with up to two million cases successfully in MLwiN.

HLM[14] (see Raudenbush *et al.*, 2001) is probably more used in North America than in the UK, but is also a powerful multilevel modelling package with a range of facilities

including data input, graphical displays, and logistic and multinomial modelling capabilities. The development of the program started as a two-level concept, although it now supports three levels. The main conceptual difference between the two packages is that, whereas MLwiN requires the model to be specified as a single entity encompassing all levels, HLM allows the user to specify the models for different levels separately. In some ways this can aid accessibility to the user and the ease with which results, including complex interactions, can be described to a non-technical audience. It is probably the case, however, that in the UK MLwiN is better understood and there is a greater community of support than for HLM. Another new package, which supports multilevel modelling in addition to a range of other modelling options is Mplus (see http://www.statmodel.com).

8.2 Acquiring suitable data

No comparability study is possible without suitable data, and for multilevel modelling to be used successfully this data needs to be comprehensive, accurate, representative and to contain all relevant variables, including the information needed to define the levels in the data (such as centre identifiers). Good data on background factors that are likely to be strongly related to outcomes (e.g. prior attainment) is essential, in order to ensure that we are comparing 'like with like'.

From the above remarks about packages, it would seem that in most cases software packages are able to cope with all available cases, so there is no merit in sampling cases for modelling – the whole dataset can be included in the analysis. However, when planning data collection it may be necessary to give consideration to sampling issues. Power calculations to determine suitable sample sizes to detect specified differences are important, but should take account of the clustering of candidates within centres and hence the design effect. Estimates of the effects of such clustering from previous studies may help to inform such calculations.

It is likely that the sampling will be done at the centre level, in which case it is important that the centres chosen are sampled randomly, probably using a stratified sampling technique that ensures they are representative in terms of important centre characteristics. If all candidates, or a fixed proportion per centre, are selected then the resulting sample will be self-weighting at the candidate level. However, if a fixed number per centre are selected, candidates in larger centres will be under-represented and consideration may need to be given to sampling with probability proportional to size to compensate for this. If there is strong clustering within centres (i.e. the centre-level variance is statistically significant compared with the candidate-level variance) then it becomes more important to get a reasonable number and spread of centres, rather than a large number of candidates per centre.

8.3 Pitfalls and problems

From all the above, some of the pitfalls and problems that may be encountered are fairly clear. These include:

- inadequate or insufficient data, or data that is biased in some way

- failure to collect suitable information, such as measures of prior attainment
- analysis that is superficial, or does not adequately model the structure of the data
- software failures, including failure to converge
- misinterpretation of the results of the modelling.

All these and other problems are controllable, but need appropriate planning, time and assistance when required. In our experience the best guard against most of these is collegiality – a community of practice that has a wide range of experience and expertise in these areas and can work as part of a team to ensure that problems do not arise or can be dealt with effectively when they do.

9 Conclusions

In this chapter we have tried to show many of the options that are available when analysing comparability study data using multilevel modelling. The challenge for the researcher is to select the appropriate model to fit the structure of the study and the type of outcome that is being modelled. Wrong choices can give misleading results, and we would strongly advise the use of variant approaches to the same data in order to obtain some idea of the sensitivity of the main results to the modelling assumptions being made.

There are big issues, discussed elsewhere in this book, which also impinge on the task of analysing comparability study data. One of the big issues is the purpose of the study. The vast majority of studies are carried out after the event, when candidates have been awarded results and the main rationale for the study is to show that results from different boards or whatever are in fact comparable. In this case, the main outcome must be to test the null hypothesis: 'Results from different boards are comparable'; if this is rejected there is no immediate action that can be taken, except to use the results to inform standard setting for the next cohort. Information on the exact degree and type of lack of comparability will be interesting, and perhaps useful for the future, but cannot directly affect outcomes.

An alternative scenario is one in which the results of a study are used to rescale results onto a common and consistent metric, for example to provide measures of school performance. Something like this happens in setting standards for National Curriculum tests, where pre-test data on the new test is used to compare with the results on the previous test and the outcomes of this analysis inform the setting of levels on the new test. In this kind of scenario the important question is not whether there is comparability, but the exact nature of the relationships determined from the model and the degree of confidence in those relationships. Details of the exact models fitted become much more important in this case, as does the need for some kind of 'sensitivity analysis'. However, even the most careful modelling requires to be interpreted in the light of the purposes of the study, the provenance of the data and the unmeasured influences that may be

operating. Analysing data of this kind is partly an art as well as a science, and no single model is likely to give us the full picture.

At this stage it is probably worth returning to the original Cresswell definition of comparability that we began with – how have our discussions in this chapter influenced our approach to this? It has to be said that the use of sophisticated modelling techniques on examining data has revealed some potential inadequacies in this supposed 'catch-all' definition. This has mainly been shown by the existence of interactions between examining boards and measures of prior attainment in some examples. It is arguable that, under Cresswell's definition, it might be possible to find two samples of pupils for which the examinations are comparable despite the significant interactions. However, at the same time it would be possible to selectively enter pupils for different boards in order to enhance their outcomes. So, are the examining boards comparable or not? We would argue that such interactions are *prima facie* evidence for lack of comparability, and a new definition is required that rules them out. We would suggest something along the lines of:

> Two examinations have comparable standards, if for all potential groups of candidates, it is not possible to selectively enter individuals for one examination or another, based on measured background information, in such a way as to improve significantly their outcomes.

Finally, what can we say about the advantages of using multilevel modelling in this kind of work? It is clear to us that the advantages of using this methodology in comparability studies far outweigh any perceived disadvantages. These advantages include:

- a unified system that encompasses other models (e.g. OLS) while allowing for hierarchical clustering
- powerful and integrated software that can fit a range of models
- the ability to allow for complicating factors, including interactions, random coefficients and complex variances, in a coherent and efficient way
- efficiency of estimation, with fewer parameters required than in alternative approaches.

Overall it is true that multilevel modelling is a powerful tool for the analysis of comparability study data, and without it the work in this area would be seriously hampered.

Endnotes

1 In principle, assuming the normality of the outcome distributions – however, the Central Limit Theorem should ensure this is a valid test for most distributions with reasonable sample sizes.

2 The so-called 'delta method' (see Eason, 1995) is an early example of a crude adjustment procedure designed to detect such relationships.

3 Although in some circumstances it may be possible to model options effects and include these in a comparability study.

4 Uniform Mark Scale (UMS) scores are based on the grade awarded and the total mark received in the examination, in such a way that the grade boundaries are defined consistently at the same UMS value. See Chapter 3 for a discussion of UMS.

5 Key Stage 3 'fine grades' are derived from scores obtained from the examinations taken, mapped on to the National Curriculum levels awarded and put on a scale such that one level = 6 points.

6 It can be argued that this within-centre or within-school homogeneity effect is a consequence of selection effects, teaching, social ordering, etc. Although it is not inevitable with educational data sets, it is sufficiently common that it should be taken into account when setting up models in this field.

7 When a correlation between two quantities is estimated using aggregated data, such as school-level mean scores, this can give a completely different result from estimating the same quantity on individuals, such as pupils. If the 'aggregated' correlation is taken as an indicator of the individual correlation, then this is described as the ecological fallacy (Robinson, 1950).

8 5.9 is the mean value of prior attainment – it is subtracted in the interaction in order to ensure it is centred about zero.

9 The logit metric allows us to model probabilities with linear functions, using the transformation $logit(x) = ln(x/(1–x))$ where $0 < x < 1$.

10 For full technical details, refer to Goldstein (2003).

11 See Chapter 9 for a fuller discussion of prior and concurrent measures.

12 See http://www.mrc-bsu.cam.ac.uk/bugs/

13 Note that the Centre for Modelling is now located at Bristol University. See http://www.cmm.bristol.ac.uk/, which also contains a review of relevant software.

14 See http://www.ssicentral.com/hlm/index.html

References

Al-Bayatti, M. (2005). *A comparability study in GCSE French. A statistical analysis of results by awarding body*. A study based on the summer 2004 examinations. London: Qualifications and Curriculum Authority.

Baird, J., & Jones, B.E. (1998). *Statistical analyses of examination standards: Better measures of the unquantifiable?* Research Report RAC/780. Assessment and Qualifications Alliance.

Bell, J.F., & Dexter, T. (2000, October). *Using multilevel models to assess the comparability of examinations*. Paper presented at the Fifth International Conference on Social Science Methodology of the Research Committee on Logic and Methodology (RC33) of the International Sociological Association, Cologne.

Bell, J.F., Vidal Rodeiro, C.L., & Malacova, E. (2004, August). *The use of multilevel logistic regression to investigate the comparability of French GCSE*. Paper presented at the Sixth International Conference on Social Science Methodology, Amsterdam.

Burstein, L. (1980). The analysis of multilevel data in education research and evaluation. *Review of Research in Education*, *8*, 158–193.

Cresswell, M.J. (1996). Defining, setting and maintaining standards in curriculum-embedded examinations: Judgemental and statistical approaches. In H. Goldstein & T. Lewis (Eds.), *Assessment: Problems, developments and statistical issues* (pp. 57-84). Chichester: John Wiley and Sons.

Dexter, T., & Massey, A. (2000, July). *Conceptual issues arising from a comparability study relating IGCSE grading standards with those of GCSE via a reference test using a multilevel model*. Paper presented at the 22nd Biennial Conference of the Society for Multivariate Analysis in the Behavioural Sciences at the London School of Economics, London.

Eason, S. (1995). *A review of the delta analysis method for comparing subject grade distributions across examining boards*. Research Report RAC/667. Guildford: Associated Examining Board.

Fitz-Gibbon, C.T., & Tymms, P. (2002). Technical and ethical issues in indicator systems: Doing things right and doing wrong things. *Education Policy Analysis Archives*, *10*(6).

Goldstein, H. (2003). *Multilevel statistical models* (3rd ed.). London: Arnold.

Gorard, S. (2003a). What is multi-level modelling for? *British Journal of Educational Studies*, *51*, 46–63.

Gorard, S. (2003b). In defence of a middle way: A reply to Plewis and Fielding. *British*

Journal of Educational Studies, *51*, 420–426.

Jones, B.E. (1997). Comparing examination standards: Is a purely statistical approach adequate? *Assessment in Education*, 4, 249–263.

Jones, B., Baird, J., & Arlett, S. (1997). *A comparability study in GCSE art and design unendorsed. A study based on the summer 1996 examinations*. Organised by the Northern Examinations and Assessment Board on behalf of the Joint Forum for the GCSE and GCE.

Mosteller, F., & Boruch, R. (2002). *Evidence matters: Randomized trials in education research*. Washington DC: Brookings Institute.

Newton, P.E. (2005). The public understanding of measurement inaccuracy. *British Educational Research Journal*, *31*, 419–442.

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 the Assessment and Qualifications Alliance (Southern Examining Group) on behalf of the Joint Forum for the GCSE and GCE.

Pinot de Moira, A. (2002a). *An inter-awarding body comparability study: The statistical analysis of results by awarding body for AS GCE and VCE business, AS GCE chemistry, AS GCE geography and AS VCE health and social care*. A study based on the summer 2001 examination and organised by the Assessment and Qualifications Alliance on behalf of the Joint Council for General Qualifications.

Pinot de Moira, A. (2002b). *Statistical robustness in comparability studies.* Unpublished report, Assessment and Qualifications Alliance.

Pinot de Moira, A. (2003). *An inter-awarding body comparability study: The statistical analysis of results by awarding body for GCE A level and AVCE business, GCE A level chemistry, GCE A level geography and AVCE health and social care*. A study based on the summer 2002 examination and organised by the Assessment and Qualifications Alliance on behalf of the Joint Council for General Qualifications.

Plewis, I., & Fielding, A. (2003). What is multi level modelling for? A critical response to Gorard (2003). *British Journal of Educational Studies*, *51*, 408–419.

Rasbash, J., Browne, W., Goldstein, H., Yang, M., Plewis, I., Healy, M., *et al.* (2002). *A user's guide to MLwiN. Version 2.1d* edition. London: University of London Institute of Education.

Rasbash, J., Steele, F., Browne, W., & Prosser, R. (2004). *A user's guide to MLwiN. Version 2.0 edition*. London: University of London Institute of Education.

Raudenbush, S., & Bryk, A. (2002). *Hierarchical linear models. Applications and data*

analysis methods (2nd ed.). Thousand Oaks, California: Sage Publications.

Raudenbush, S., Bryk, A., Cheong, Y., & Congdon, R. (2001). *HLM 5: Hierarchical linear and nonlinear modeling*. Lincolnwood, Illinois: Scientific Software International, Inc.

Robinson, W. (1950). Ecological correlations and the behaviour of individuals. *American Sociological Review*, *15*, 351–357.

Stobart, G., Bibby, T., & Goldstein, H., with Schagen, I., & Treadaway, M. (2005). *Moving to two-tier GCSE examinations: An independent evaluation of the 2005 GCSE pilot and trial*. London: Qualifications and Curriculum Authority.

Styles, B. (2006). Educational research v. scientific research. *Research Intelligence*, *95*, 7–9.

Tymms, P., & Vincent, L. (1995). *Comparing examination boards and syllabuses at A-level: Students' grades, attitudes and perceptions of classroom processes.* Curriculum, Evaluation and Management Centre, University of Newcastle-upon-Tyne. Technical report commissioned by the GCE Examining Boards. Belfast: Northern Ireland Council for the Curriculum, Examinations and Assessment.

While, D., & Fowles, D. (2000). *A comparability study in GCSE mathematics. Statistical analysis of results by board. A study based on the work of candidates in the summer 1998 examinations.* Organised by the Assessment and Qualifications Alliance (Northern Examinations and Assessment Board) on behalf of the Joint Forum for the GCSE and GCE.

Yang, M., Goldstein, H., Browne, W., & Woodhouse, G. (2002). Multivariate multilevel analyses of examination results. *Journal of the Royal Statistical Society A*, *165*(1), 137–153.

COMMENTARY ON CHAPTER 10

Anne Pinot de Moira

Ian Schagen and Dougal Hutchison open their chapter by correctly warning against the assumption that statistical modelling provides objectivity. They describe model-fitting as an art rather than a science where 'there is no substitute for experience and a deep understanding of the subject matter'. While this is indisputably true, the element of subjectivity in any statistical analysis extends beyond the choice of model, the formulation of dependent variable, the decision to include given independent variables and the sampling of data. Even for a technically sound model, the findings are only valid to the extent they are interpreted legitimately.

In the literature a naïve faith in statistical significance testing is blamed for creating the illusion of much sought-after objectivity in research work (Schmidt, 1996). The system where a null hypothesis is defined and then rejected whenever the probability of being wrong in that decision is less than some critical value is appealing in the sense that it is rule-based. In his paper of 1951, Yates lamented the emphasis placed upon tests of significance suggesting they are often regarded as the 'ultimate objective'. As Tukey (1991) observed, however, a null hypothesis is always false at some level of decimal places; adding an element of futility to such an objective. Cohen (1990) made the same argument more forcefully:

> If [the null hypothesis] is false, even to a tiny degree, it must be the case that a large enough sample will produce a significant result and lead to its rejection. So if the null hypothesis is always false, what is the big deal about rejecting it?

The abundance of candidate-level data for national examinations in England brings Cohen's observations into sharp focus when considering the interpretation of comparability. Carver (1978) writes that 'statistical significance ordinarily depends on how many subjects are used in the research'. The larger the sample size, the smaller the difference between effects which will be detected as statistically significant. In the context of national examinations, where data are plentiful, it seems sensible that the modelling and comparison of standards should not rely solely on statistical significance testing. It has long been argued that the magnitude of the difference between effects, or effect size, is of much greater practical significance (Cohen, 1988; Kirk, 1996).

Consider a study of GCSE English which used a sample of data to compare the grading standards applied between awarding bodies (Pinot de Moira, 2000). A logistic multilevel model was fitted to data to determine the probability of exceeding the foundation tier grade C threshold dependent upon awarding body of entry. There were 6,651 candidates nested within 111 centres. The resultant model suggested no overall statistically significant difference in grading dependent upon awarding body (Table 1).

Of real interest to the educational practitioner, however, should be the magnitude of the difference in grading standards between awarding bodies.

Table 1 Estimates for the two-level logistic model describing the log odds of a GCSE English candidate exceeding the foundation tier grade C threshold

					Joint	
		β	se	p	χ^2	p
Fixed Effects	Constant	1.245	0.287	0.000		
	English Key Stage 3 result	1.008	0.062	0.000		
	Mean mathematics & science Key Stage 3 result	0.188	0.069	0.007		
	Mean GCSE result	1.296	0.061	0.000		
	Female	0.668	0.084	0.000		
	Awarding Body 2	–0.914	0.348	0.009	8.821	0.066
	Awarding Body 3	–0.558	0.344	0.105		
	Awarding Body 4	–0.783	0.322	0.015		
	Awarding Body 5	–0.433	0.345	0.210		
Random Effects Centre level		0.658	0.119	0.000		

(Awarding body 1 is set as the base category.)

In their introduction to multilevel modelling, Snijders and Bosker (1999) describe effect size as an approximate relationship between the standard error of an effect, the power of the test and the significance level (Equation 1).

$$\text{Effect Size } (\gamma) \approx (z_{1-\alpha} + z_{1-\beta}) \times \text{se } (\gamma) \qquad (1)$$

Where $z_{1-\alpha}$ is a z score associated with the significance level of α and $z_{1-\beta,}$ is the z score associated with a given power $1-\beta$[1]. The z scores are derived from the standard normal distribution. For the purposes of inter-awarding comparability, let us define:

$$\text{Effect Size } (\gamma) = \gamma_A - \gamma_B$$

Where γ_A is the parameter estimate for awarding body A which is greater than γ_B the parameter estimate for awarding body B.

In the current context, therefore, effect size is described as the difference between two awarding bodies in the log odds of exceeding a given grade threshold. Using the model displayed in Table 1 for illustration, the standard error associated with the awarding body effects can be estimated as approximately 0.4. From Equation 1 the effect size which would be detected at a significance level $\alpha = 0.05$ and power $1-\beta = 0.8$ would be approximately 0.996. Such a statistic has little useful meaning but, because the model is logistic, the effect size can be transformed to be expressed in terms of a probability. Expressed as a difference in the probability of exceeding a grade threshold, effect size becomes a statistic with practical utility.

$$\text{Effect Size } (\gamma) = \gamma_A - \gamma_B$$

$$= \ln\left(\frac{p+\delta}{1-(p+\delta)}\right) - \ln\left(\frac{p}{1-p}\right)$$

Where
- p is the probability of exceeding a given grade threshold
- δ is the difference in probability of exceeding a given grade threshold between awarding bodies B and A
- se (γ) is now defined as an approximation of the standard error associated with the awarding body parameter estimates

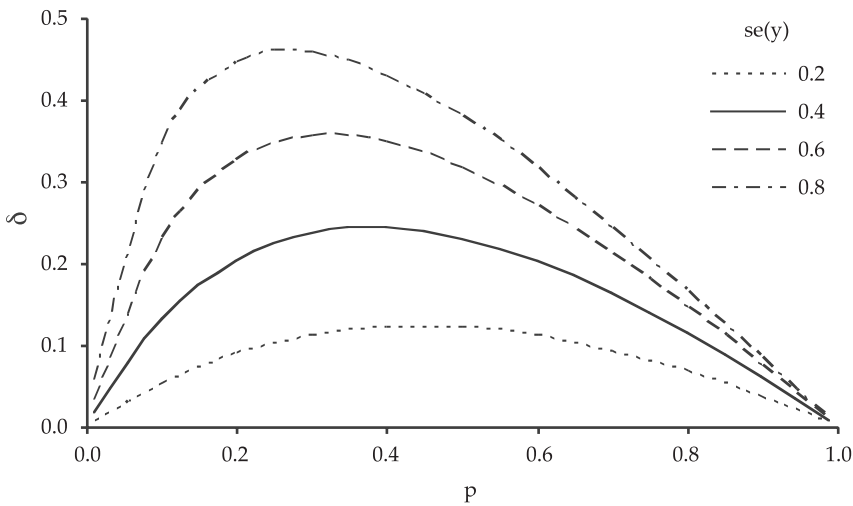
The minimum difference between awarding bodies that would be detected as statistically significant is estimated thus,

$$\delta \approx \frac{p - p^2 - e^m + p^2 e^m}{p - 1 - pe^m}$$

Where $m = (z_{1-\alpha} + z_{1-\beta}) \times se(\gamma)$

For the GCSE English multilevel model (Table 1), the minimum detectable difference between awarding bodies would be described by the solid curved line in Figure 1. Figure 1 also describes the relationship between δ and p for other values of se (γ) where α=0.05 and 1–β=0.8.

Figure 1 The relationship between δ and p for a logistic model with varying values of se(γ) where the significance level a=0.05 and power 1–β=0.8



Among the candidates included in the multilevel model just over 20% were awarded a grade C. An initial estimate of the probability of exceeding the grade C threshold would therefore be 0.2 with the minimum statistically detectable difference between awarding bodies also being 0.2 (see Figure 1). While the statistical significance tests

flagged no overall differences between the awarding bodies (Table 1), the subsequent analysis of effect size suggests that differences of up to 20% in the award of grade C between award bodies would be regarded as statistically non-significant. However, a difference of such magnitude would clearly be unacceptable, not least because the specifications and the awarding processes would lack face validity and, therefore, lose credibility in the eyes of the public.

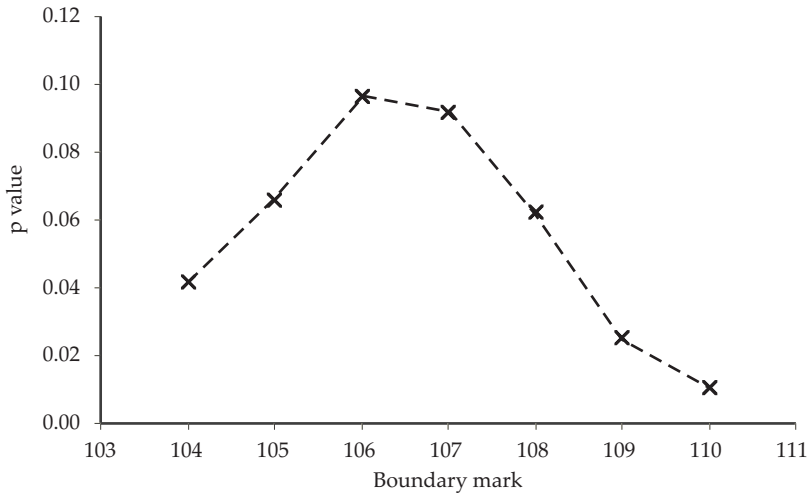
The process of evaluating effect size allows a translation of the statistical outcomes from a complex model into terms to which those involved in standard-setting can relate. Even though an effect size expressed as a probability provides more information about comparability than an unqualified statistical test, standards in national examinations are set by determining boundary marks for each of the contributing components. An effective assessment of comparability between awarding bodies would require reference to the mark scale in order to fulfil a remit of advising awarding committees of any necessary remedial action.

The extent to which comparability between national examinations can be expected is, however, limited by the fact that most are marked on a discrete ordinal scale and marking is completed before determination of grade boundaries. It would therefore be unrealistic to expect that grading standards could be *exactly* the same between awarding bodies (an interesting observation when considered alongside Tukey's (1991) assertion that a null hypothesis is always false at some level of decimal places). 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 statistically significantly lenient to statistically significantly severe. Delap (1992) discussed this matter in the context of grade award meetings and the maintenance of year-on-year grading standards.

Empirically, it is possible to explore the statistical sensitivity of grading standards to a mark scale which is discrete. Returning to the GCSE English data, the grade C boundary for Awarding Body 5 was determined as 105 in the award meeting. Figure 2 illustrates the effect that repositioning this boundary would have had on the joint statistical test applied to assess comparability of the GCSE English specifications. Given a naïve assumption that comparability is assured if the p-value is more than 0.05, a grade boundary placed in the range 105–108 would lead to the conclusion that grading standards are aligned. These 'satisfactory' extremes of grade C boundary would award between 31.3% and 25.8% of candidates a grade C.

Rather than accepting that grading standards are aligned with grade boundaries in the range 105–108, a judgement is required as to whether the extremes exceeding the grade threshold are defensible. When setting grade boundaries to maintain year-on-year comparability some awarding bodies have derived acceptable deviances, in percentage terms, between years. All other things being equal, for large-entry subjects, the percentage of candidates exceeding a given grade threshold is not expected to vary from the previous year by more than 2%. Further work would be needed to establish whether the use of such acceptable deviances could be extended to between-

Figure 2 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 awarding body specification



awarding body comparability. However, it is contextual information of this nature which should be fed into the design and interpretation of a model of comparability. Indeed a retrospective look at effect size might never be needed were an educational researcher more often afforded the luxury of a controlled experiment where data could be sampled to target a particular hypothesis and power calculations could be performed in advance of any analysis. Instead, inter-awarding body comparability studies are largely based on opportunity samples with missing data and self-selected entry patterns. It is essential, therefore, that the limitations of any data are explored both before and after analysis and that, to be of practical value, model outcomes are related to the measurement scale where remedial action can be effected.

As Ian Schagen and Dougal Hutchison correctly conclude 'the challenge for the researcher is to select the appropriate model to fit the structure of the study and the type of outcome that is being modelled'. Implicit within this challenge must be the understanding that outcome should be presented in terms that are relevant to the target audience (Schagen, 2004). Consequently, when discussing the findings from an ordered categorical multilevel model (Table 5), for example, the authors helpfully transform their findings to an 'intuitive metric'.

When considering comparability in the national examinations of England, the researcher must go beyond presenting the headline news, which in itself might be misleading, to suggest to awarding committees appropriate remedial action. Furthermore, it should be recognised that the cocktail of large datasets and small mark ranges, available for national comparability studies, makes blind acceptance of

statistical hypothesis testing a risky business. The researcher would do well to heed the advice given by Reese (2004):

> Calculating statistical significance is a tool, a step in the process of analysis. The interpretation of a result requires the researcher's knowledge, in particular to put new data into the context of previous scientific knowledge.

The modelling techniques presented in Chapter 10, and indeed throughout the rest of the book, provide powerful instruments with which to describe data but, without valid interpretation and contextualisation, the statistics produced are utterly redundant in a practical sense.

Endnote

1 The significance level of a statistical test is the probability (α) of wrongly rejecting the null hypothesis if it is actually true (Type I error). The power of a statistical test is the probability ($1-\beta$) that it will correctly reject the null hypothesis if it is actually false.

References

Carver, P.C. (1978). The case against statistical significance testing. *Harvard Educational Review, 48*(3), 378–399.

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences* (2nd ed.). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, *45*(12), 1304–1312.

Delap, M.R. (1992). *Statistical information at awarding meetings: The discrete nature of mark distributions*. Research Report RAC/585. Guildford: Associated Examining Board.

Kirk, R.E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement*, *58*(5), 746–759.

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 the Assessment and Qualifications Alliance (Southern Examining Group) on behalf of the Joint Forum for the GCSE and GCE.

Reese, R.A. (2004). Does significance matter? *Significance*, *1*(1), 39–40.

Schagen, I. (2004). Presenting the results of complex models – Normalised coefficients, star wars plots and other ideas. In I. Schagen & K. Elliot (Eds.), *But what does it mean? The use of effect sizes in educational research* (pp. 25–41). Slough: National Foundation for Educational Research.

Schmidt, F.L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods, 1*(2), 115–129.

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

Tukey, J.W. (1991). The philosophy of multiple comparisons. *Statistical Science, 6*(1), 100–116.

Yates, F. (1951). The influence of statistical methods for research workers on the development of the science of statistics. *Journal of the American Statistical Association*, *46*(253), 19–34.

COMMENTARY ON CHAPTER 10

Peter Tymms

There can be no doubt that Hierarchical Linear Models (aka multilevel models, MLMs) represent a major advance and that they have solved some key statistical problems through the use of clever algorithms and modern computing power. They have provided useful new perspectives and they rightly continue to be at the forefront of some aspects of educational research. But they have not been accepted without question; see, for example, Fitz-Gibbon (1997); De Leeuw and Kreft (1995); Gorard (2003a; 2003b; 2004; 2007). There have also been staunch defences; see, for example, Plewis and Fielding (2003). Space does not permit a full discussion but five broad issues are set out below to indicate some key points.

When should multilevel models be used?

A number of researchers have carried out studies in which they have analysed data using traditional methods only to find that in order to get published they have been required to use multilevel models. When doing as they were told and getting their papers published, they discovered that the newly analysed results hardly differed from the original.

Schagen and Hutchison state that because Ordinary Least Squares (OLS) approaches assume that the units are independent the OLS results will give 'biased estimates of the statistical significance'. This must be true, but that is not the point. The question is 'how big is the bias?' This was extensively investigated as part of The Value Added National Project (see, for example, Trower & Vincent, 1995) and the findings summarised in Fitz-Gibbon (1997). Not only were the regression coefficients identical to all intents and purposes but the OLS and MLM 'school effects' were found to correlate around 0.99 with one another. This held when the analyses involved both linear or curvilinear relationships and single or multiple predictors. But, it is argued, a primary advantage of MLMs is that the errors are better estimated than in OLS regression. Again this must be true but how big is the difference? Tymms (in Fitz-Gibbon, 1997, p. 110) found 'little difference' between the errors on the coefficients in OLS regressions and MLMs. He also found almost identical estimates of errors on the residuals from the two procedures. This was not the case in the example shown by Schagen and Hutchison. Under what circumstances are important differences found?

Given a choice between a simple approach and a more complex analysis it could be argued that one should always go for an MLM since, as Schagen and Hutchison note, that is the standard against which other analyses will be judged. But there is the matter of communication. They state 'we believe that the objections to the difficulty of the technique are overstated'. Have they successfully explained variance at the second level to a politician or a journalist?

They also write 'it is more important to be correct than to appear simple'. But the issue may not be a choice between such stark alternatives and it seems reasonable to ask what amount of clustering ensures that traditional approaches are misleading. Clearly the lower the intra-class correlation the less the importance of explicitly modelling the clustering; the key question is about the degree to which the results of analyses might mislead. Indeed, Kennedy and Mandeville (2000) state 'the question of when to use multilevel modelling is an important one.'

Complexity

Can a model be too complex? Assuming that we do have situations where the clustering is large enough to mean that it makes sense to use MLMs, the models themselves can become exceedingly complex. There can be several levels with cross-classification, interactions, and any or all variables used at the lower levels can reappear at a higher level in some aggregate form. Such models can become extremely difficult to construct, understand and interpret. But is it possible for a model to become too complex? The well-respected statistician Steve Raudenbush (1994) advises on the need for 'parsimonious pre-specified models'. He does this because of the 'precariousness of knowledge based on exploratory analyses using trimmed models and retrospective explanation'. Not only are MLMs statements of a theoretical position but they also tend to be isolated and self-referential. This leads to the question 'when does the model become so complex it becomes impossible to gainsay?' This asks about the falsifiability of the chosen model which is in some senses a theory. In the same way that Popper (1963) writes about the falsifiability of scientific theories, we should ask about the falsifiability of multilevel models.

Terminology

Within the reports of MLMs the writing often involves words as such as 'effects', 'explanation' and 'impact'. Schagen and Hutchison use the phrases, 'the model explains the data', 'Board C effect' and 'do have an impact on'. These are misleading terminologies as they all imply causal relationship. Occasionally, very occasionally, MLMs are used to analyse the data from randomised control trials (see, for example, Tymms and Merrell, 2006), but most commonly MLMs deal with passive observational data. As Kennedy and Mandeville (2000) note 'for purposes of drawing inferences about school effects students would be randomly assigned to schools and schools would be randomly assigned to process conditions'. Naturally, the sophisticated users of MLMs are aware of the issue but they do little to discourage the use of the established terminology.

How can the causally-laden words often used in statistics be discouraged?

Errors on predictors

In OLS regression and in MLMs, except where dummy variables are used, there are errors on the predictors, and yet the errors are assumed to be non-existent. This can lead to problems. If, for example, attainment were modelled using socio-economic status as a predictor at the pupil level and at another level, perhaps the board level, using the average socio-economic status, then it is quite likely that a significant

compositional effect would be found. But that compositional effect would inevitably appear because of the error of measurement in the predictor. Indeed the extent to which the predictor is measured with error can be used to indicate the extent to which a compositional effect will appear. This phantom effect is a consequence of one of the assumptions of multilevel models. The difficulty was established before the advent of MLMs (Hauser, 1970) but has been demonstrated more recently using MLwiN (Harker and Tymms, 2004).

Bias from shrinkage

When constructing MLMs there is a danger that unstable relationships can appear at the second, or higher, levels because very small units can produce wild results. To get round this the results at the higher levels are shrunk in proportion to the reliability of their measurement. This produces more stable models; however, the shrinking introduces bias. A very small unit with a genuinely high or low value is artificially shrunk towards the mean, hiding its true colours. Comparability studies will not usually need to worry about this issue but it could arise if small samples are being analysed or when the standards adopted for a less popular syllabus from one awarding body are being assessed.

References

De Leeuw, J., & Kreft, I.G.G. (1995). Questioning multilevel models. In I.G.G. Kreft (Ed.), Hierarchical linear models: Problems and prospects [Special issue]. *Journal of Educational and Behavioral Statistics, 20*(2), 171-189.

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

Gorard, S. (2003a). What is multi-level modelling for? *British Journal of Educational Studies*, *51*, 46-63.

Gorard, S. (2003b). In defence of a middle way: A reply to Plewis and Fielding. *British Journal of Educational Studies*, *51*, 420-426.

Gorard, S. (2004). Comments on modelling segregation. *Oxford Review of Education, 30*, 435-440.

Gorard, S. (2007). The dubious benefits of multi-level modelling. *International Journal of Research and Method in Education*, *30*(2), 221-236.

Harker, R., & Tymms, P. (2004). The effects of student composition on school outcomes. *School Effectiveness and School Improvement*, *15*(2), 177-199.

Hauser, R.M. (1970). Context and consex: A cautionary tale. *American Journal of Sociology*, *75*, 645-664.

Kennedy, E., & Mandeville, G. (2000). Some methodological issues in school effectiveness research. In C. Teddlie & D. Reynolds (Eds.), *The international handbook of school effectiveness research* (pp.189-205). London: Falmer Press.

Plewis, I., & Fielding, A. (2003). What is multi level modelling for? A critical response to Gorard (2003). *British Journal of Educational Studies*, *51*, 408-419.

Popper, K.R. (1963). *Conjectures and refutations: The growth of scientific knowledge*. London: Routledge & Kegan Paul.

Raudenbush, S. (1994). Searching for a balance between a priori and post hoc model specification: Is a 'general approach' desirable? *School Effectiveness and School Improvement*, *5*(2), 196-198.

Trower, P., & Vincent, L. (1995). *The value added national project technical report*. London: School Curriculum and Assessment Agency.

Tymms, P., & Merrell, C. (2006). The impact of screening and advice on inattentive, hyperactive and impulsive children. *European Journal of Special Needs Education*, *21*(3), 321-337.

RESPONSE TO COMMENTARIES ON CHAPTER 10

Ian Schagen and Dougal Hutchison

Response to Anne Pinot de Moira

To a large extent we agree with the comments about statistical significance. It is not enough for something to be statistically significant – the size of the effect is important. See, for example, Schagen and Elliot (2004) for a discussion of effect sizes in educational research. However, effect sizes and significance tests are not mutually incompatible: the latter attempts to answer the question 'is there evidence for a difference due to something other than chance?' while the former addresses the question: 'how big is the difference and does it matter?' Both are important in comparability studies, and can be addressed through multilevel analysis.

The comment about using 'statistically significant' rather than just 'significant' is probably good practice. However it can turn out very cumbersome, and it is often more elegant to say, for example, 'significantly different' than 'statistically significantly different'. Given that our chapter is very technical, we feel that it is reasonable to assume that readers will not confuse the technical use of the word with the common English usage.

We agree with the final comment about the need for valid interpretation and contextualisation, alongside powerful and appropriate modelling tools and high quality data.

Response to Peter Tymms

Peter Tymms takes a more general overview of the value and application of multilevel modelling, and queries some of our justifications for advocating its use in the comparison of examining boards. Most of his comments are unexceptionable but some criticisms are more general in their focus than the current application. In many ways we are inclined to feel that the confrontation between Ordinary Least Squares (OLS) or logistic regression on the one hand and multilevel modelling (MLM) is more apparent than real. In fact, it is possible to consider OLS to be a special case of MLM, with zero variance at higher levels, in much the same way as a strictly hierarchical model could be considered a special case of a model with nested and crossed variance components. Each has its place, and can provide valuable information, and one has to balance statistical soundness, ease of use, and computing feasibility.

A cautious statistician would usually want to test the hypothesis that higher level variances were zero before proceeding further, which would mean running the MLM anyway – and then why not proceed in this way, especially if the hypothesis is rejected? Tymms' convention that MLM is significantly more complex to run than OLS is not really true with modern software. MLM is certainly more powerful and can fit more complex models in situations where they are needed.

It is generally accepted that the fixed coefficients in a variance component model will be close to those in an OLS model, though they will not generally be completely the same. However one finds that the standard errors of fixed coefficients in MLM, especially those of variables at higher levels, will typically be larger, and in some cases substantially larger, than those of OLS models, and thus coefficients are less likely to be statistically significant. In a sense in this situation therefore it is less a question of MLM finding new effects, but rather of MLM not finding effects that are not there.

Modelling is never an exact science, but a balancing act between parsimony and fitting the data well. The chapter is not intended to present a precise recipe for carrying out comparability studies, but to show what models are possible and the consequences of omitting certain elements from the model, for example higher level variances or random slopes. Communicating the results of any sort of model may well be a challenge, but key elements should not be omitted because of this. Simple formulations can often be found to explain, for example, higher level variances: for example, '84% of the variation in the results was between pupils in the same schools, and 16% due to differences between schools'.

As Tymms points out, at some value of intra-class correlation the OLS and MLM models are likely to give very similar results. What this value is will vary from case to case. Even if we knew this critical value, we would then need to run an analysis to determine the intra-class correlation for our data before deciding to opt for OLS or MLM. Busy statisticians such as ourselves prefer to eliminate this extra work and go into MLM which will estimate all the required parameters and give the same results as OLS if the intra-class correlation is effectively zero.

To a large extent, choice between examining boards at the level of individual subjects is a school-level decision. In this connection Anne Pinot de Moira's illustration of an inter-board comparison in her comment on our paper is highly relevant. One's first reaction is that a sample of over 6,000 is going to be big enough to identify any phenomenon that is actually of any real world importance. However, since the board decision is more nearly a school-level one than an individual one, it turns out that using MLM probably makes a substantial difference. She has very kindly agreed to re-run the analyses from her commentary, using a non-hierarchical model, and the relevant part of the results is shown in Table 1.

Table 1 Non-hierarchical modelling of data from commentary by Pinot de Moira

	MLM Analysis			**OLS Analysis**		
Exam board effects	β	**Se**	χ^2	β	**Se**	χ^2
Awarding Body 2	–0.914	0.348		–0.558	0.136	
Awarding Body 3	–0.558	0.344	8.821	–0.304	0.132	35.427
Awarding Body 4	–0.783	0.322	(P=0.066)	–0.558	0.120	(P=0.000)
Awarding Body 5	–0.433	0.345		–0.147	0.125	

Columns 2 and 3 correspond to the results in the table in her commentary, run using MLM, and 5 and 6 are the corresponding results using a non-hierarchical model. It can be seen that there are some differences in the β coefficients for the awarding bodies between the two analyses. More important in this context however is the difference in size of the standard errors of these coefficients, which are very much larger in the MLM than in the OLS analysis. The results in the OLS analysis are thus highly significant statistically, while those in the more appropriate MLM are not.

Our exposition of the benefits of MLM did not confine itself to simple fixed inter-board differences. An important aspect of inter-board differences lies in the fact that they vary between subgroups, between schools and over ability ranges, and this formed an important part of the examining boards' contention that there was no 'quick fix' of apparent differences between subjects (see Newton, 1997, for a discussion of this point). We believe that MLM is the appropriate statistical technique for addressing questions of this type.

Tymms also mentions that there is little difference between the estimated standard errors for school residuals using OLS and MLM, which may or may not be true, but does not appear relevant to our exposition. He also comments that apparent aggregated school-level effects may be due to a biasing effect of measurement error. This is certainly true (see Hutchison, 2007, for an extensive exposition) but again is not particularly relevant. Finally he refers to the discussion on whether shrunken residuals should be used for second or higher level units (for example, in these applications, schools or examination centres) – again, an interesting point for discussion, but not relevant here. These last three points indeed appear to have strayed in from some other paper.

Finally we confess that in writing the chapter we did on occasions use terminology which could be interpreted causally, and this is something which ideally should be avoided. It is not always possible to find circumlocutions which avoid this without becoming clunky and interrupting the flow of an argument, but this is clearly an area where we need to improve our language and the way we use it.

References

Hutchison, D. (2007). When is a compositional effect not a compositional effect? *Quality and Quantity*, *41*, 219–232.

Newton, P.E. (1997). Measuring the comparability of standards between subjects: Why our statistical techniques do not make the grade. *British Educational Research Journal*, *23*, 433–449.

Schagen, I., & Elliot, K., (Eds.). (2004). *But what does it mean? The use of effect sizes in educational research*. Slough: National Foundation for Educational Research.