School Effectiveness and School Improvement
1997, Vol. 8, No. 4, pp. 369–395

0924-3453/97/0804-0369$12.00

Methods in School Effectiveness Research*

Harvey Goldstein
Institute of Education, University of London

ABSTRACT

This paper discusses the methodological requirements for valid inferences from school effectiveness research studies. The requirements include long term longitudinal data and proper statistical modelling of hierarchical data structures. The paper outlines the appropriate multilevel statistical models and shows how these can model the complexities of school, class and student level data.

INTRODUCTION

The term 'school effectiveness' has come to be used to describe educational research concerned with exploring differences within and between schools. Its principal aim is to obtain knowledge about relationships between 'explanatory' and 'outcome' factors using appropriate models. In its basic form it involves choosing an outcome, such as examination achievement, and then studying average differences among schools after adjusting for any relevant factors such as the intake achievements of the students. Researchers are interested in such things as the relative size of school differences and the extent to which other factors, such as student social background or curriculum organisation, may explain differences. All of this activity is set within the context of the well known relationship between intake characteristics and outcomes and the fact that schools do not acquire students at random.

*I am most grateful to the following for helpful comments on a draft. John Gray, Kate Myers, Ken Rowe, Pam Sammons, Jaap Scheerens, Louise Stoll, and Sally Thomas.

Correspondence: Harvey Goldstein, Institute of Education, University of London, 20 Bedford Way, London, WC1H 0AL, UK. E-mail: h.goldstein@ioe.ac.uk.

Manuscript submitted: May, 1996
Accepted for publication: April 21, 1997

The earliest influential research was that of Coleman et al. (1966), followed by Jencks et al. (1972), both being based around traditional multiple regression techniques. These studies included a very large number of schools and school level measurements as well as various measures of student socio-economic background. They were not longitudinal, however, and so were unable to make any intake adjustments. There followed the influential work of Rutter, Maughan, Mortimore, Ouston, and Smith (1979) which *was* longitudinal, but inconclusive since it involved only 12 schools. The first study to incorporate the minimum requirements necessary for any kind of valid inference was the Junior School Project (JSP) (Mortimore, Sammons, Stoll, Lewis, & Ecob, 1988). It had a longitudinal design, sampled 50 schools and used multilevel analysis techniques. Since that study there have been important developments in the design, the conceptualisation and the statistical models available for school effectiveness research. This paper describes the current state of methodological understanding and the insights which it can provide. It does not directly address the issue of choosing appropriate outcomes nor intake measures. Clearly, such choices are crucial if we wish to make causal inferences, but they are beyond the scope of the current paper, which is methodological.

PERFORMANCE INDICATORS AND SCHOOL COMPARISONS

During the 1980s and early 1990s, in the UK and elsewhere, considerable attention was given by school effectiveness researchers to the production and use of 'performance indicators', usually measures of average school achievement scores. A considerable debate developed, often driven by political considerations, over the appropriateness or otherwise of using achievement (and other) output measures for ranking schools, and this has extended to other kinds of institutions such as hospitals (Goldstein & Spiegelhalter, 1996; Riley & Nuttall, 1994).

The difficulties associated with performance indicators are now well recognised and are twofold. First, their use tends to be very narrowly focused on the task of *ranking* schools rather than on that of establishing factors which could *explain* school differences and secondly, a number of studies have now demonstrated that there are serious and *inherent* limitations to the usefulness of such performance indicators for providing reliable judgements about institutions (Goldstein & Thomas, 1996). Briefly, the reasons for these limitations are as follows.

First, given what is known about differential school effectiveness (see below) it is not possible to provide simple, one- (or even two-) dimen-

sional summaries which capture all of the important features of institutions. Secondly, by the time information from a particular institution has been analysed, it refers to a 'cohort' of students who entered that institution several years previously so that its usefulness for *future* students may be dubious. Even where information is analysed on a yearly basis, for reasons which will become clear, it is typically necessary to make adjustments which go back two or more years in time. Furthermore, it is increasingly recognised that institutions, or teachers within those institutions, should be judged not by a single 'cohort' of students, but rather on their performance over time. This makes the historical nature of judgements an even more acute problem.

It is now well understood that institutional comparison has to be based upon suitable adjustments for intake achievement and other relevant factors, but even when this can be done the resulting 'value added' estimates usually have too much uncertainty attached to them to provide reliable rankings. This point will be illustrated in a later section, and is particularly important when comparisons are based upon individual subject departments where the number of students may be small. In addition there is always the difficulty that the statistical model we are using may fail to incorporate all the appropriate adjustments, or in some other way may be misspecified. At best, value added estimates can be used as crude screening devices to identify 'outliers' (which might form the basis for follow-up research), but they cannot be used as definitive statements about the effect of a school per se (Goldstein & Spiegelhalter, 1996; Goldstein & Thomas, 1996). Thus we may be able to establish that differences exist among schools, but we cannot, with any useful precision, decide how well a *particular* school or department is performing: this 'uncertainty principle' operates to provide a fundamental barrier to such knowledge. The same set of problems occurs when studying changes in value added estimates over time in order to judge 'improvement' (Gray, Jesson, Goldstein, Hedger, & Rasbash, 1995).

For a similar reason, schemes which attempt to provide individual schools with value added feedback *for their own use* often have dubious validity. Typically there is too much uncertainty associated with both choice of model and relatively small student numbers, especially when considering individual classrooms or subjects. The efficacy of such schemes as the principal basis for judging effectiveness has little evidential support, and there appears to be a paucity of detailed and independent evaluation of existing enterprises such as the British ALIS project (FitzGibbon, 1992) or the Tennessee Value Added System (Sanders & Horn, 1994). It should also be remembered that any estimates obtained for indi-

vidual institutions are relative ones; that is they position each institution in relation to the other institutions with which they are being compared. If the comparison group is not representative of the population of interest, for example because it is self selected, then we may have some difficulty in interpreting the individual estimates. By the same token, it is perfectly possible for *all* schools to be performing satisfactorily in some absolute sense while still exhibiting differences. The use of the descriptions 'effective' and 'ineffective' therefore may be quite misleading unless this is understood, and it would be more accurate to qualify such descriptions by the term 'relative' whenever they are used.

Despite these reservations, the use of adjusted school or classroom estimates to detect very discrepant units does have certain uses. As a device for Education Authorities or others to indicate where further investigations may be useful it seems worth pursuing: but this is a *higher level* monitoring function carried out on groups of institutions as a screening instrument. If handled with care, such data may also be useful as a component of schools' own self evaluation (Bosker & Scheerens, 1995).

I have no wish to deny that individual schools should be held accountable through the collection of a wide range of relevant information: my point is that little understanding is obtained by attempting to do this, *principally and essentially indirectly*, through simple indicators based upon student performance.

THE EMPIRICAL FRAMEWORK FOR SCHOOL EFFECTIVENESS RESEARCH

In later sections I will explore the statistical models used in school effectiveness research, but it is useful first to look at some conceptual models in order to establish a framework for thinking about the issues.

Experimental Manipulation

In a scientifically ideal world we would study matters of causation in education by carrying out a succession of randomised experiments where we assigned individuals to institutions and 'treatments' at random and observed their responses and performances. We would wish randomly to assign any chosen student, teacher, and school factors a few at a time to judge their effects and thus, slowly, hope to discover which forms of organisation, curriculum, classroom composition, etc. were associated with desirable outcomes. Naturally, in the real world we cannot do this, but it is instructive to imagine what we *would* do within such a programme and then

decide how closely we can approach it by using the observations, measurements and statistical tools which do happen to be available.

The research questions are very broad. To begin with, there will be several 'outcomes' of interest such as different kinds of academic performance or aspects of the 'quality' of the school experience (see, for example, Myers, 1995). There are also questions about how to measure or assess such outcomes. We may be interested not merely in outcomes at the end of one stage of schooling, but multiple outcomes at each of several stages, for example at the end of each school year if we wish to study teacher effects. More generally we may wish to establish a dynamic model of change or progression through time where the whole period of schooling is studied across phases and stages. In order to focus the discussion, and without sacrificing too much generality, I shall raise briefly a series of questions about 'effectiveness', which, also for convenience, I shall take to be judged by academic performance, although, as I have pointed out, issues about which measurements to adopt are extremely important.

The model which has informed most school effectiveness work to date consists of a set of schools, measurements of outcomes on students, and other measurements on the schools and their staff. Suppose that we are at liberty to take students ready for entry to Primary school, assign them at random among a suitably large sample of schools, follow them until they leave and measure their achievements at that point. We are at liberty to assign class and head teachers at random: we can select them by age, gender or experience and then deposit them across the schools in a systematic fashion. We can vary such things as class size, curriculum content, school organisation, and school composition, and if we wish can even designate selected children to change schools at particular times. With sufficient time and resources we can then observe the relationships between outcomes and these design factors and so arrive at a 'causal' understanding of what appears to matter.

The term 'appears' is used advisedly: however well planned our study there still may be important factors we have omitted and we need to remember that our studies are always carried out in the (recent) past and hence may not provide sure guides to the future. Key issues therefore are those of stability and replicability. Is the system that we are studying stable over time so that what exhibits a 'causal' relationship now will continue to do so in the future? If not, then we need to extend our understanding to predict *why* relationships at one time become modified. This requires a *theory* about the way such relationships are modified by external social conditions since generally we cannot subject changes in society to experimental manipulation.

Undoubtedly, there are some issues where experimentation will be more successful than others. In the area of programme evaluation, the random assignment to different 'treatments' is perhaps the only safe procedure for establishing what really works, and often there will be a rationale or theoretical justification for different programmes. On the other hand, for example in studies of the effect of class size upon achievement, experimental manipulation seems to be of limited usefulness: it may be able to establish the existence of overall effects but may then be faced with the larger problem of explaining what, for example, it might be about teaching that produces such effects.

Statistical Adjustment

On returning to the real world of schooling the common situation is that we have no control over which children attend which schools or are assigned to which teachers. More generally we have little control over how individual teachers teach or how classes are composed. The best we can do is to try to understand what factors *might* be responsible for assigning children to schools or teachers to ways of organising teaching. Outside the area of programme evaluation, a large part of school effectiveness research can be viewed as an attempt to do precisely this and to devise satisfactory ways of measuring such factors.

It is well established that intake characteristics such as children's initial achievements and social backgrounds differ among schools for geographical, social and educational reasons. If we could measure accurately all the dimensions along which children differed it would be possible, at least in principle, to adjust for these simultaneously within a statistical model so that schools could be compared, notionally, given a 'typical' child at intake. In this sense we would be able to measure 'progress' made during schooling and attribute school differences to the influence of schools per se.

In practice, however, even if we could make all the relevant measurements, it is unlikely that the adjustment formula would be simple. Complex 'interactions' are possible, so that children with particular combinations of characteristics may behave 'atypically' and extremely large samples would be needed to study such patterns. Furthermore, events which occur during the period of schooling being studied may be highly influential and thus should be incorporated into any model. Raudenbush and Willms (1995) refer to the process of carrying out adjustments for *initial* status as an attempt to establish 'type A' comparisons between institutions and they point out that such comparisons are those which might be of interest to people choosing institutions, although as I have already

pointed out, such choices are inherently constrained. The task facing school effectiveness research is to try to establish which factors are relevant in the sense that they differ between schools and also that they may be causally associated with the outcomes being measured. In this respect most existing research is limited, contenting itself with one or two measures of academic achievement and a small number of measures of social and other background variables, with little attempt to measure dynamically evolving factors during schooling. There is a further difficulty which has been highlighted in recent research. This is that the measurement of student achievement at the start of a stage of schooling cannot estimate the *rate of progress* such children are making at that time and it is such progress that may also affect selection into different schools as well as subsequent outcome. More generally, the entire previous achievement, social history and circumstances of a student may be relevant, and measurements taken at a single time point are inadequate. From a data collection standpoint this raises severe practical problems since it requires measurements to be made on children over very long periods of time. I shall elaborate upon this point later.

In addition to measurements made on students it is also relevant to suppose that there are further factors which may influence progress. Environmental, community and contingent historical events may alter that progress. Intake and developmental factors may interact with such external factors and with the characteristics of schools. Thus, for example, students with low intake achievement may perform relatively better in schools where most of their fellow students have higher as opposed to lower achievements: girls may perform better in single-sex schools than in mixed schools or family mobility during schooling may affect performance and behaviour. Together with school process measurements these factors need to be taken into account if we wish to make useful, causally connected, inferences about the effects of schools on progress, what Raudenbush and Willms (1995) refer to as 'type B' effects. We also need to be prepared to encounter subtle interactions between all the factors we measure: to see whether, for example, girls in girls' schools who are low achievers at intake and in small classes with students who tend to be higher achievers, perform substantially better than one would predict from a simple 'additive' model involving these factors.

Additional to this framework is the need to replicate studies across time and place, across educational systems, and with different kinds of students. In the face of such complexity, and given the practical difficulties of data collection and the long time scale required, progress in understanding cannot be expected to take place rapidly. It is, therefore, impor-

tant that the planning of school effectiveness studies is viewed from a long term perspective and that available resources are utilised efficiently. The following discussion is intended to show how statistical modelling can be used in order both to structure the data analysis and to provide a suitable framework for long term planning.

INFERENCES FROM EXISTING QUANTITATIVE RESEARCH INTO SCHOOL EFFECTIVENESS

In a comprehensive review of school effectiveness research, Scheerens (1992) lists a number of factors, such as 'firm leadership' and 'high expectations', which existing research studies claim are associated with 'effective' schooling. His view (Chapter 6) is that only 'structured teaching' and 'effective learning time' have received adequate empirical support as factors associated with effectiveness. Similarly, Rowe, Hill, and Holmes-Smith (1995) emphasise that current policy initiatives are poorly supported by the available evidence, and that clear messages are yet to emerge from school effectiveness research. Two of the same authors (Hill & Rowe, 1996) also point out that inadequate attention has generally been paid to the choice and quality of outcome measures. The views of these authors', together with the fact that very few studies, if any, satisfy the *minimum* conditions for satisfactory inference, suggest that few positive conclusions can be derived from existing evidence. The minimum conditions can be summarised as:

(1) that a study is longitudinal so that pre-existing student differences and subsequent contingent events among institutions can be taken into account;
(2) that a proper multilevel analysis is undertaken so that statistical inferences are valid and in particular that 'differential effectiveness' is explored;
(3) that some replication over time and space is undertaken to support replicability;
(4) that some plausible explanation of the *process* whereby schools become effective is available.

This is not to criticise all existing studies. Many of these studies have helped to clarify the requirements I have listed. Nor do I wish to argue that we should refrain from adopting *policies* based upon the best available evidence, from research and elsewhere. Rather, my aim is to set out current and future possibilities and I shall do so by describing a suitable

framework for data modelling and analysis, and the issues which need to be addressed.

STATISTICAL MODELS FOR SCHOOL EFFECTIVENESS STUDIES

The standard procedure for deriving information about relationships among measurements is to model those relationships statistically. This section will develop such models, elaborating them where necessary without undue statistical formalism. A more detailed technical description can be found, for example, in Goldstein (1995).

Measurement

A word about 'measurement' is appropriate. Although most of what I have to say is in the context of cognitive or academic achievement, it does in principle apply to other attributes, such as attitudes, attendance, etc. My use of the term 'measurement' is intended to be quite general. It refers not just to measures such as test scores made on a continuous or pseudo-continuous scale, but also to judgmental measures concerning, say, mastery of a topic or attitude towards schooling. All of these kinds of measurements can be handled by the models I shall discuss or by straightforward modifications to them. For simplicity, however, I deal largely with the case of a continuous outcome or 'response' measurement.

There are many issues to be resolved in devising useful measures: most importantly they must have acceptable validity (suitably defined) and they must be replicable. I shall not discuss these requirements in any more detail, except to point out that no matter how good any statistical model may be, if the measurements are poor or inappropriate then any conclusions will be suspect.

Single Level Models: Using Student Level Data Only

The original work of Coleman (Coleman et al., 1966), Jencks (Jencks et al., 1972) and Rutter (Rutter et al., 1979) was about relationships among student level variables, but ignored the actual ways in which students were allocated to schools. This results in two problems. The first is that the resulting statistical inferences, for example significance tests, are biased and typically over-optimistic. The second is that the failure explicitly to incorporate schools in the statistical model means that very little can be said about the influence of schools per se. What is required are models which simultaneously can model student level relationships and

take account of the way students are grouped into individual schools. In the next section I describe some simple models of this kind.

MULTILEVEL MODELS

It is now generally accepted that a satisfactory approach to school effectiveness modelling requires the deployment of multilevel analysis techniques. The classic exposition of this, together with a detailed discussion of some of the difficulties, is in the paper by Aitkin and Longford (1986). Multilevel modelling is now an established technique with a growing body of applications, some of it highly technical (see Goldstein, 1995, for a detailed review). Nevertheless, its basic ideas can be expressed in simple statistical terms, and in view of the centrality of this technique in school effectiveness research I shall take a few paragraphs in order to convey the essential components.

The simplest realistic multilevel model relates an 'outcome' or 'response variable' to membership of different institutions. For convenience, suppose we are dealing with Primary schools and have a measure of reading attainment at the end of Primary school on a random sample of students from each of a random sample of Primary schools.

If y_{ij} is the reading score on the i-th student in the j-th school we can write the following simple model

$$\begin{aligned} y_{ij} &= \beta_j + e_{ij} \\ &= \beta_0 + u_j + e_{ij} \end{aligned} \tag{1}$$

which says that the reading score can be broken down into a school contribution (β_j) and a deviation (e_{ij}) for each student from their school's contribution. In the second line we have decomposed the school contribution into an overall mean (β_0) and a departure from that mean for each school. These departures (u_j) are referred to as school 'residuals'.

So far (1) is unremarkable, merely re-expressing the response, our reading test score, into the sum of contributions from students and schools. In traditional statistical terms this model has the form of a one-way analysis of variance, but as we shall see it differs in some important respects. Our first interest lies in whether there are any differences among schools. Since we are treating our schools as a random sample of schools in order to make generalisations about schools at large, we need to treat the u_j as having a distribution among schools. Typically we assume that this distribution is Normal with a zero mean (since we have already accounted for

the overall population mean by fitting β_0) and variance, say σ_u^2. The student 'residual' e_{ij} is also assumed to have a variance, say σ_e^2.

The first question of interest is to study the size of σ_u^2. If, relative to the total variation, this is small then we might conclude that schools had little effect, or putting it another way, knowing which school a student attended does not predict their reading score very well. (We shall see later that such a judgement on the basis of a simple model like (1) may be premature.)

The total variation is simply

$$\text{var}(y_{ij} - \beta_0) = \text{var}(u_j + e_{ij}) = \sigma_u^2 + \sigma_e^2 \tag{2}$$

since we assume that the u_j, e_{ij} vary independently, and we define the 'intra-school correlation' as

$$\sigma_u^2 / (\sigma_u^2 + \sigma_e^2)$$

which measures the relative size of the between-school variance and also happens to be equal to the correlation of reading scores between two students in the same school.

We can 'fit' such a model by taking a data set with students identified by the schools they belong to and then estimating the required parameter values ($\beta_0, \sigma_u^2, \sigma_e^2$). This can be accomplished using different software packages, the most common ones being VARCL (Longford, 1987), HLM (Bryk & Raudenbush, 1992) and MLn (Rasbash & Woodhouse, 1995). For some of the more complex models discussed later the first two packages are too limited. The models can also be fitted by the BUGS package based on Gibbs Sampling (Gilks, Richardson, & Spiegelhalter, 1996). It should be noted that the most common statistical packages used by social scientists have very limited procedures for multilevel analysis, although this situation undoubtedly will change.

Estimating Residuals

In addition to estimating the variation between schools we may also be interested in the individual values of the residuals u_j, usually interpreted as the 'effect' associated with each school. The first thing to note is that the accuracy of any estimates we can make of these quantities will depend largely on the number of students in each school. Secondly, there are essentially two ways in which we might obtain the estimates. The simplest procedure would be to calculate the mean for each school and then subtract the *overall* mean from each one of these to obtain the school

residuals. This, in effect, is what we would obtain from a traditional one way analysis of variance applied to (1). If we have large numbers of students in each school this will provide reasonable estimates. Where, however, a school has a very small number of students, sampling variations imply that the mean will not only be poorly estimated (have a large confidence interval) but may also turn out by chance to be very large or very small. It is for this latter reason that an alternative procedure is usually preferred. The resulting estimates are referred to as 'shrunken' residuals, since in general they will usually have a smaller variation than the true school means (as estimated by σ_u^2). They can be motivated in the following manner.

Consider the prediction of an unknown u_j from the set of observed scores $\{y_{ij}\}$ in the j-th school. This school may be one of the ones used in the analysis or it may, subsequently, be a new school. In practice we base the prediction on the differences between the observed scores and the *fixed part* prediction of the model, in this case just β_0. This can be viewed as a multiple regression having the following form

$$u_j = \alpha_1(y_{1j} - \beta_0) + \alpha_2(y_{2j} - \beta_0) + ... + \alpha_{n_j}(y_{n_j j} - \beta_0) \tag{3}$$

where the regression coefficients $\{\alpha_k\}$ are derived from the *random* parameters of the model, that is they depend on the quantities σ_u^2, σ_e^2. In this simple 'variance components' model the required estimate turns out to be

$$\hat{u}_j = \frac{n_j \sigma_u^2}{n_j \sigma_u^2 + \sigma_e^2} \tilde{y}_j, \qquad \tilde{y} = \frac{\sum_i (y_{ij} - \beta_0)}{n_j} \tag{4}$$

which is the estimate from our first simple procedure ($\tilde{y}$) multiplied by a shrinkage factor which always lies between zero and one. As n_j (the number of students in the j-th school) increases and also as σ_e^2 increases relative to σ_u^2 this factor approaches one, and the two estimators tend to coincide. For small n_j the estimate given by (4) moves towards zero so that the estimated school mean (the residual added to the overall mean β_0) moves towards the overall (population) mean. The shrinkage estimates therefore are 'conservative', in the sense that where there is little information in any one school (i.e., few students) the estimate is close to the average over all schools. It is the assumption in model (1) that the school residuals belong to a distribution (whose parameters we can estimate), which results in schools with few students having estimates near to the mean of this distribution. In the extreme case where we have no information on the students in a school, our best estimate is just this overall mean. In

addition, of course, each shrunken estimate will also have a sampling error enabling us to place a confidence interval about it to measure the uncertainty.

To illustrate these points I shall use data from the JSP study (Mortimore et al., 1988). These consist of 728 students in 48 Junior schools with the response or outcome measure being a mathematics test score measured on the students at the age of 11 years. Table 1 gives the parameter estimates and Figure 1 plots the 48 residual estimates ordered by value. The error bars are the estimated 95 % confidence intervals about each estimate. The variance of these estimated residuals is 3.53 which is just 68 % of the estimate in Table 1, illustrating the operation of the shrinkage factor. The intra-school correlation of 12 % is within the range of estimates from similar studies.

An important feature of Figure 1 is the extent to which the confidence intervals for each school cover a large part of the total range of the estimates themselves, with three quarters of the intervals overlapping the population mean (zero). This is a feature which is found also in more complex models where intake and other factors have been taken into account. It illustrates that attempts to rank or separate individual schools in league tables may be subject to considerable uncertainty (see Goldstein & Spiegelhalter, 1996). The extent of this uncertainty is determined by the intra-school correlation and the number of students available for analysis. Comparisons of institutions should therefore always provide estimates of uncertainty, as in Figure 1. I shall not pursue this issue further, except to remark that residuals also have technical uses for the purposes of making judgements about model adequacy (Goldstein, 1995).

Table 1. Variance Components Model for 11 Year Old Mathematics Scores: JSP Data.

Fixed Part	Estimate	Standard error
β_0	30.61	0.40
Random Part		
Between schools: σ_u^2	5.16	1.55
Between students: σ_e^2	39.3	1.9
Intra school correlation	0.12	

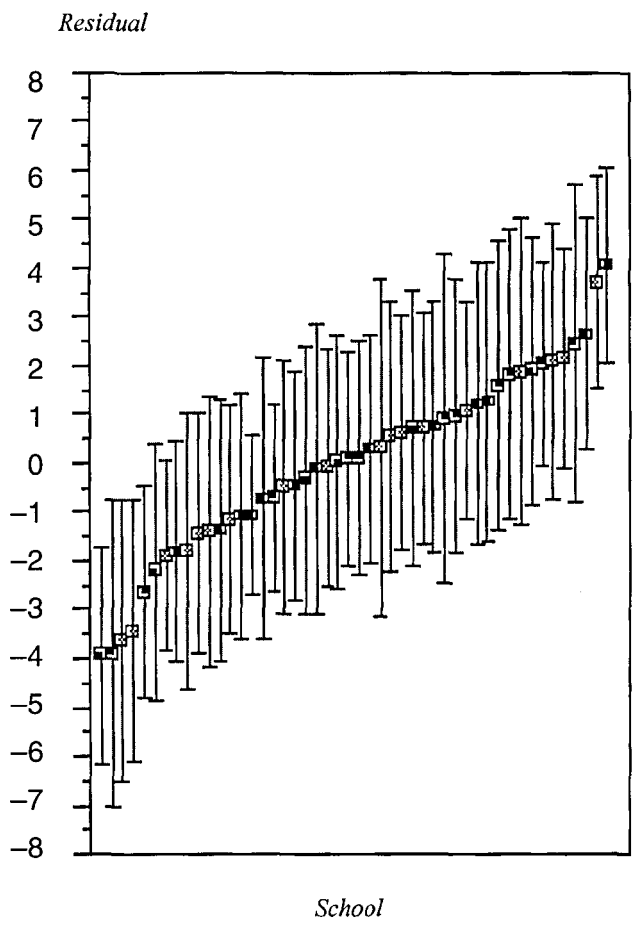

Fig. 1. Ordered school residuals with 95 % confidence intervals for model in Table 1.

CONTEXTUALISATION AND ADJUSTMENT IN STATISTICAL MODELS

It is generally found that the most powerful predictor of achievement at the end of a period of schooling, is the achievement measured at the start of the period. For example, various studies quote correlations for the phase of secondary schooling (a period of some five years) as between 0.5 and 0.7 (Scheerens, 1992). In terms of other outcomes, such as student attitudes or behaviour there has been far less work on identifying appropriate 'intake' factors, although Mortimore and colleagues (1988)

studied these outcomes and Myers & Goldstein (1996) explore some of the practicalities of carrying out adjustments for such measures.

The analysis of Table 1 is now elaborated to include as an adjustment variable a Mathematics test taken at the age of eight years by the same children at the start of the Junior school period. The statistical model now becomes

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + u_j + e_{ij} \tag{5}$$

where x_{ij} is the eight year score and β_1 represents the average predicted increase in eleven year Mathematics achievement for a unit increase in eight year score.

The between-school and the between-student variances are smaller than before, with the intra-school correlation remaining virtually unchanged. Model (5) implies that the relationship between intake and outcome is the same for each school with the residuals u_j now representing the difference between the mean eleven year score predicted for each school at any given eight year score and that predicted for the population as a whole. These adjusted residuals are often known as 'value added' school estimates because, by adjusting for initial achievement they attempt to measure the relative progress made in different institutions – what those institutions can be considered to have 'added' to the initial scores. This term, however, is somewhat difficult to justify. The intake and output measurements are normally different and since the comparison is always a relative one, the connotation of 'adding value' seems somewhat misleading. The term 'adjusted comparison' is more accurate.

We can now add other factors or variables to (5), such as socio-economic status, gender, etc. When this is done we find that there is a negli-

Table 2. Variance Components Model for 11 Year Old Mathematics Scores, Adjusting for Eight Year Mathematics Attainment Measured about its Mean: JSP Data.

Fixed part	Estimate	Standard error
β_0	30.6	0.35
β_1	0.61	0.03
Random Part		
Between schools: σ_u^2	4.03	1.18
Between students: σ_e^2	28.1	1.37
Intra school correlation	0.13	

gible effect due to gender, with boys and girls performing at about the same level after adjusting for intake, whereas for social class there is an advantage of 1.3 units on the eleven year score scale to those from non-manual (white collar) families: in other words, from a given starting point the non-manual children exhibit more progress. There is also some evidence that the relationship with the eight year Mathematics score is 'less steep' for the non-manual children. Such models have become the standard ones used in school effectiveness studies. If we wish to investigate the role of factors such as 'leadership', 'time on task', etc., we would include suitable measurements on these into (5) in just the way we have introduced gender and social class, remembering that many factors may change during the course of schooling.

Differential Effectiveness

The most important further elaboration to model (5) is to allow the relationship between outcome and intake to vary from school to school, sometimes known as the differential effectiveness model. Formally, (5) now becomes

$$\begin{aligned} y_{ij} &= \beta_0 + \beta_{1j} x_{ij} + u_j + e_{ij} \\ \beta_{1j} &= \beta_1 + v_j \end{aligned} \tag{6}$$

so that the 'slope' term β_{1j}, has a subscript, j, indicating the school with which it is associated, and there is an overall population mean slope of β_1. This implies that there are two random variables at the school level, each with a variance and in general a non-zero correlation. In the present case, fitting (6) yields variances of 4.61 and 0.034 for the 'intercept' and slope with $\beta_1 = 0.61$ as before. Since the eight year score has been centred at its mean the intercept variance is the between-school variance at the mean eight year score value. The standard deviation of the slope across schools is therefore $\sqrt{0.034} = 0.18$ which implies a considerable variation between the smallest and largest slopes. To illustrate this Figure 2 shows the predictions of eleven year score for three randomly chosen schools.

What is most noticeable here is that for high intake achievements there is little difference between schools, whereas for low intake scores there is a much bigger 'effect'. This particular example is pursued further in Woodhouse (1996). In that analysis the data are modelled in more detail, including the allocation of different variances for boys and girls and the transformation of the test scores so that they have more nearly Normal distributions. Although it is not pursued here, the issue of transformations

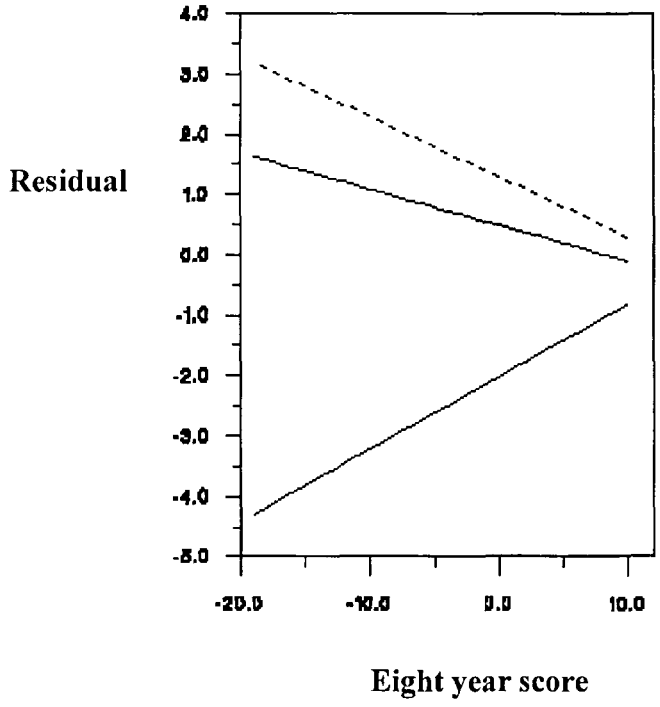

Fig. 2. Predicted eleven year scores for three schools by eight year score.

is an important one. Since very many of the measurement scales used in education are somewhat arbitrary, it is reasonable to consider transforming them in order to simplify the statistical modelling. In the present case a Normal score transformation reduces the apparent differential effectiveness and generally it needs to be remembered that any substantive interpretations are with respect to the particular measurement scale used and that a (non-linear order preserving) transformation of that scale may lead to somewhat different interpretations. Woodhouse (1996) even gives an example where a Normal score transformation removes a random coefficient from a model and by so doing also removes the interpretation of differential effectiveness!

The differential effectiveness model, nevertheless, seems to be widely applicable in school effectiveness work. It implies that schools need to be differentiated along more than one dimension. For example, schools which appear to be 'effective' for low achieving students are not necessarily the same as those which are 'effective' for high achieving students. In the remaining sections I discuss important elaborations of these basic models which it is important to incorporate into future work.

In addition to adjusting for student level variables, there is considerable interest in what are often termed 'contextual' effects, namely variables operating at the level of the classroom or school. One kind of contextual effect is a characteristic of a school or teacher, for example teacher age or size of school. Another kind refers to an aggregate characteristic of the student body, such as the average intake achievement or the average socio economic status. Of particular interest is the possibility of interactions between individual and group-level characteristics. Also, in addition to the average values of such characteristics we may find that, say, the spread of intake achievement scores within a classroom is an important predictor.

Finally, there is the important, and largely ignored, problem of measurement errors. It is well known that where a predictor variable, such as an intake test score, has a low reliability, then inferences from a statistical model which ignores this can be seriously biased. In multilevel models this can affect estimates of both the fixed coefficients, such as class or gender differences, and also the estimates of between-unit variances (see Goldstein, 1995, Chapter 10). Work is currently proceeding on methods for the efficient handling of this problem (Woodhouse, Yang, Goldstein, Rasbash, & Pan, 1996).

USING AGGREGATE LEVEL DATA ONLY

Many early attempts to model relationships between achievement outcomes and school factors relied on the use of average achievement scores for each school together with measurements, say, of average socio-economic status or average intake test scores (see for example Marks, Cox, & Pomian-Srzednicki, 1983). The main difficulty with such analyses is that they tell us nothing about the effects upon individual students. Relationships at the school level may be very different from those found at the student level. In addition, the former do not allow us to study whether relationships are the same for different kinds of students, whether they vary from school to school, or how well student achievement can be predicted from a knowledge of intake achievement and other factors. As first pointed out by Robinson (1950), we can draw seriously misleading inferences about individual relationships from aggregate data. Woodhouse & Goldstein (1989) illustrate this with examination data and show also that estimates of aggregate level relationships can be very unstable.

To illustrate the possibly misleading nature of such inferences I have carried out two analyses using data from 66 Inner London schools and

5562 students (see Goldstein et al., 1993, for a detailed description). The response variable is a score based upon the 16 year old school leaving examination, the GCSE. A reading test score at the start of secondary school (the London Reading test) is available as are a number of school characteristics. For present illustrative purposes I have used only the gender composition of the school (boys only, girls only, or mixed) together with the reading test score and the student's gender as predictors or explanatory variables. Table 3 presents the results of four analyses. The first is an aggregate level analysis using school average values only. The second is corresponding multilevel 'variance components' analysis of the kind described above, fitting a 'between-student' and a 'between-school' variance. The third adds the student level gender to the model and the fourth is a differential effectiveness model where the relationship between the exam score and reading test score varies from school to school. The reading test score and GCSE score have been transformed, using Normal scores, to have Normal distributions.

In Table 3 the coefficient estimates for the school level variables, the contrasts between boys schools and mixed schools, and between girls schools and mixed schools are similar in the two analyses A and B, whereas the coefficients for the student level variable, the reading test score, are very different. If we estimate school level residuals from each

Table 3. Multilevel versus Aggregate Level Models with Aggregate Level Predictors of GCSE.

	Estimate (SE)			
	A	B	C	D
Fixed				
Intercept	–0.02	–0.03	–0.09	–0.05
LRT score	1.08 (0.09)	0.52 (0.01)	0.52 (0.01)	0.52 (0.02)
Girls - Mixed school	0.17 (0.06)	0.18 (0.07)	0.10 (0.08)	0.17 (0.06)
Boys - Mixed school	0.05 (0.08)	0.03 (0.09)	0.09 (0.10)	0.04 (0.08)
Girls - Boys (student level)	–	–	0.14 (0.03)	–
Random				
Between-student variance	–	0.56 (0.01)	0.56 (0.01)	0.55 (0.01)
Between-school variance	0.05 (0.01)	0.07 (0.01)	0.07 (0.01)	0.07 (0.01)
LRT coefficient variance	–	–	–	0.01 (0.003)
LRT/Intercept covariance	–	–	–	0.02 (0.006)

Note. Analysis A fits the aggregate level model, analysis B the multilevel model, analysis C adds the student level gender to the multilevel model and analysis D adds LRT as a random coefficient at the school level to model B.

analysis we find that the correlation between them is just 0.79. In analysis C we elaborate the model to include the student's own gender, and we see how this changes the inferences for the effect of the school type, by adjusting, *at the student level*, for the better performance of girls and showing that the apparently superior effect of girls schools is simply a result of misspecification of the model by using only aggregate level predictors. We could of course pursue the analysis of gender effects further by studying the effect of the proportion of girls in mixed schools.

Another major difficulty with aggregate level analyses is their inability to model random coefficients, that is to study differential effectiveness. To illustrate the difference this can make even with simple models we have refitted model B in Table 3, adding a random coefficient for the LRT at school level. The estimates are given in analysis D, and we see a substantial between-school variation among the LRT coefficients, with a correlation of 0.76 between the intercept, that is the school effect at the mean LRT score, and the LRT coefficient. As in Figure 2 this implies different relationships among schools depending on the LRT score of the students. For illustration I have chosen a student who has an LRT score at the 95th percentile and calculated the predicted residual for each school,

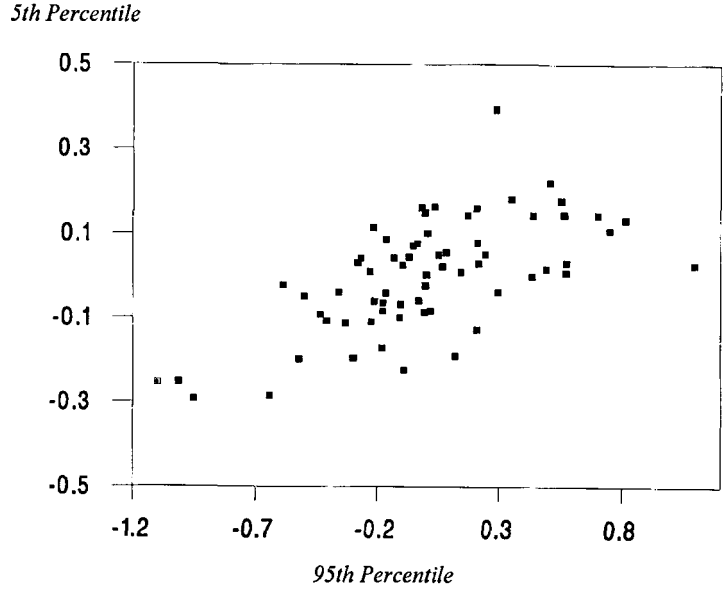

Fig. 3. Estimated school effects for students at 5th and 95th LRT scores.

and then compared these with the residuals for a student at the 5th percentile. The results are given in Figure 3 which is a scatterplot of these two estimates, which have a correlation of 0.64. As is clear, there is a considerable difference among schools in the expectations for students with high and low reading scores on entry. Analyses which ignore such differences, and in particular judgements about schools based on only aggregate level analyses which cannot take account of such differential effects, are not merely incomplete, but potentially very misleading.

CROSS CLASSIFICATIONS

Educational systems are only partially organised into hierarchies, with students grouped within schools, schools within education authorities, etc. If we follow students from, say, Primary into Secondary schools, then each student will belong to a particular Primary and a particular Secondary school (ignoring for simplicity the students who change Primary or Secondary schools). In other words, if we *cross classify* the Primary schools by the Secondary schools, each cell of that classification will contain students from a particular Primary and a particular Secondary school. Not all the cells need have students in them; some Primary/Secondary combinations may not be present. In addition we may be able further to classify students by the neighbourhood where they live in. The interesting research question is how far neighbourhood and membership of the two kinds of schools influences our outcomes. To illustrate the importance of taking account of cross classified structures, I consider the JSP data set where the same students were followed up after Junior school into Secondary school and their 16 year GCSE examination results recorded. I examine the extent to which these examination scores are influenced by the Junior as well as the Secondary school attended. Full details of the analysis can be found in Goldstein & Sammons (1997).

The statistical model now becomes the following elaboration of (5), and of course we can further elaborate with random coefficients and other predictor or explanatory variables.

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + u_{j(J)} + u_{k(S)} + e_{ij} \tag{7}$$

In (7) $u_{j(J)}$ is the contribution from the j-th Junior school, and $u_{k(S)}$ is the contribution from the k-th Secondary school. When we fit the model we estimate a between-Junior school variance and a between-Secondary school variance (the Junior and Secondary school effects are assumed to operate

independently). Goldstein & Sammons (1997) found that the Junior school variation was about three times as large as the Secondary school variation. In part, this is simply because Secondary schools are larger and can be regarded as averaging scores over a number of Junior schools. Nevertheless, the analysis also found that those Junior schools associated with high scores at eleven years of age when students transferred to Secondary school, were also those Junior schools whose students tended to perform better at GCSE. The continuing influence of Junior school attended seems clear.

One conclusion from these results is that when studying school differences it is important to pay attention to the institution attended in a prior phase of schooling. To adjust for achievement (or other factors) only at a single point of time when students are changing schools is to ignore the fact that students will be developing at different rates. Including previous institutional membership in the model will partly account for this but in general we need to be able to take a *series* of measurements on students prior to starting any phase of schooling in order properly to account for intake characteristics, and of course we would also wish to be able to measure factors during that phase of schooling. This implies very long term data collection strategies and these in turn will require the modelling of more complex cross classified structures because of factors such as mobility (see below).

CLASSROOM LEVEL ANALYSES

Rowe et al. (1995), as well as other authors, point out that since learning takes place within classrooms as well as within schools, the classroom level is a key one to model. If this is done it will usually imply that measurements are taken at least at the start and end of each school year when students enter and leave classes. In many educational systems students, at least during the Primary or Elementary stage, progress from year to year in intact groups. In this case we do not have a cross classification, say for year 2 against year 1, since all the year 2 students in a class were in the same class in year 1. In other systems, however, classes are effectively resorted after each year and a full cross classified analysis has to be carried out to ascertain the contributions from each class and teacher in each year.

Among others, Hill and Rowe (1996) find that when classrooms are included as a level between the student and the school, the between-classroom variation in achievement is larger than that between schools and the latter is often reduced to a very small value indeed. The class-

rooms studied by Hill and Rowe were in Primary unstreamed schools and their conclusion is that school effectiveness research needs to be located nearer to the classroom than the school level. In studying classrooms we also need to address the issue discussed above, namely that comparisons among classrooms should adjust for prior achievements at more than one previous time point. It will also be necessary to monitor carefully the way in which students are actually allocated to classes and teachers since this may be connected in subtle ways with achievement progress.

In secondary schools self-contained classrooms may not exist, with students being allocated to different teachers for different subjects, etc. Consider first the case where, say, ten different subjects are on offer, and students are obliged to take three compulsory core subjects and three others. Suppose also that for each subject there is only a single class. We now have what is referred to as a 'multivariate' data structure where there are six responses for each student and not the same set for each student. As in the case of a single outcome measure, we will be interested in the factors associated with each response and between-school differences. An example is where we wish to study between-school variation in examination results for subject departments. In addition, there will be interesting information contained in the correlation, at both the student and school level, between the different subject responses. For example, Goldstein et al. (1993) found a rather low correlation at the school level between Mathematics and English examination results at sixteen years among London children after adjusting for intake. This provides, therefore, another kind of dimension along which schools may differ.

In the second case, suppose that for some of the subjects, perhaps just the compulsory ones, more than one class is formed. We could now add a between-classroom level, for these subjects, to our multivariate model. Now, however, we will also have a potential cross classification at the classroom level since at least for the three compulsory subjects students will be subject to a 3-way classification by the class attended. It is easy to see how a great deal of complexity can be built up. Indeed, such complexity will *usually* be present and any claim fully to represent the processes of schooling needs to recognise this in the data analysis.

MOVING BETWEEN SCHOOLS

During any phase of schooling many students will change their institution. In some systems the turnover may be very large indeed, for example accounting for most students during Primary schooling. Existing longitu-

dinal school effectiveness studies almost invariably analyse only those students who remain in the same institution for the duration of the study. Yet this is problematical since those who move may well have different characteristics and the effectiveness of a school, it could be argued, may partly be judged in terms of such turnover. Certainly, if in a school it is those who make less progress who leave more often, then any estimates based only on those who stay will be biased. This raises two kinds of difficult issues. The first is that of data collection. To deal with the problem any study will need to be able to track those who move, and will need to be able to measure them within their new institutions. Likewise, those students who enter a study institution after the study has begun will need to be included and measured after entry.

There is also a problem for the statistical modelling. Those students who enter an institution after the beginning of a study will have data missing, for example on intake measurements. Procedures for dealing with such 'missing' data have been developed, and Goldstein (1995, Chapter 11) describes a way of handling this problem for multilevel models. For students who move to new institutions we may have all the individual measurements, but now have to consider the fact that they 'belong' to more than one institution and Goldstein (1995, Chapter 8) also indicates how such models may be specified.

A related issue is the one where a student is known to belong to a single higher level unit but it is not known which one. Such a situation may occur in the case of classrooms where the identification information for some students has been lost, or it may be that in a cross-classified analysis of Secondary by Primary schools, some of the Primary school identifications may be missing. In such circumstances we will normally be able to assign a *probability* of belonging to each potential unit (possibly equal probabilities in the case of complete ignorance) and this information can then be used in the data analysis, without losing the students' data. Hill and Goldstein (in press) discuss this and related issues with examples.

The problems discussed are practical ones which arise with real data and for which there are now techniques available. Future school effectiveness studies need to find ways of ensuring that they are designed so that these issues can be addressed properly.

IMPLICATIONS

It will be apparent from this discussion of existing school effectiveness research and the models which are available, that there is a considerable potential for new and important understandings. The statistical models now available, together with powerful and flexible software, enable researchers properly to explore the inherently complex structures of schooling in a manner that begins to match that complexity. At the same time there are inescapable political implications of research into school effectiveness which need to be recognised and which interact with the methodology.

The role and functioning of educational institutions is of major public concern in many educational systems, and any research which aims to explain why such institutions vary will inevitably create interest. In the case of 'league tables' such interest is often far from benign and, as I have indicated, much school effectiveness research has been side-tracked into a rather fruitless activity around the production and refinement of such rankings. Advances in the ability to process and model large amounts of student and school data have assisted this process, and it has sometimes seemed that this has been justified by the mere application of sophisticated techniques. Several authors have become worried about this and in particular the exploitation of school effectiveness research for purely political ends. Hamilton (1996) and Elliot (1996) have each expressed concerns of this kind as well as more general criticisms of the direction taken by some school effectiveness researchers. Sammons, Mortimore, and Hillman (1996), and Sammons and Reynolds (1996) have responded to these criticisms, emphasising the positive contributions that school effectiveness studies can make.

There is little doubt that politicians of almost all persuasions have seized upon some of the published interpretations of school effectiveness research in order to promote their own concerns. Examples of such concerns are a desire to introduce a competitive marketplace in education, and a wish to locate the responsibility for both success and failure with schools, while ignoring the social and environmental context in which schools operate (Goldstein & Myers, 1996; Myers & Goldstein, 1996). Yet, as I have attempted to demonstrate, much of the existing research is methodologically weak, so that strong conclusions about why schools become effective or why they 'fail' are difficult to sustain.

In my view, little of the true potential of school effectiveness research yet has been realised. Even the minimum requirements for valid inference are demanding ones. They imply considerable expenditure of thought and

resources as well as the long term follow-up of large numbers of students, schools and classrooms. They need to be replicated across educational systems, phases and types of schooling. It follows that those engaged in school effectiveness research need to convey this modesty of knowledge when they make claims to understanding.

REFERENCES

Aitkin, M., & Longford, N. (1986). Statistical modelling in school effectiveness studies. *Journal of the Royal Statistical Society, A, 149*, 1–43.

Bosker, R. J., & J. Scheerens (1995). A self evaluation procedure for schools using multilevel modelling. *Tijdschrift voor Onderwijsresearch*, 20, 154–64.

Bryk, A. S., & S.W. Raudenbush (1992). *Hierarchical linear models*. Newbury Park, CA: Sage.

Coleman, J., Campbell, E., Hobson, C., McPartland, J., Mood, A., Weinfield, F., & York, R. (1966). *Equality of educational opportunity*. Washington, DC: US Government Printing Office.

Elliot, J. (1996). School effectiveness research and its critics: Alternative visions of schooling. *Cambridge Journal of Education, 26*, 199–223.

FitzGibbon, C. T. (1992). *School effects at A level: Genesis of an information system?* In D. Reynolds & P. Cuttance (Eds.), *School effectiveness, research policy and practice*. London: Cassell.

Gilks, W. R., Richardson, S., & Spiegelhalter, D.J. (1996). *Markov chain Monte Carlo in practice*. London: Chapman and Hall.

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

Goldstein, H., & Myers, K. (1996). Freedom of information: towards a code of ethics for performance indicators. *Research Intelligence, 57*, 12–16.

Goldstein, H., Rasbash, J., Yang, M., Woodhouse, G., Pan, H., Nuttall, D., & Thomas, S. (1993). A multilevel analysis of school examination results. *Oxford Review of Education, 19*, 425–433.

Goldstein, H., & Sammons, P. (1997). The influence of secondary and junior schools on sixteen year examination performance, a cross classified multilevel analysis. *School Effectiveness and School Improvement, 8*, 219–230.

Goldstein, H., & Spiegelhalter, D. (1996). League tables and their limitations: Statistical issues in comparisons of institutional performance. With discussion. *Journal of the Royal Statistical Society, A, 159*, 385–443.

Goldstein, H., & Thomas, S. (1996). Using examination results as indicators of school and college performance. *Journal of the Royal Statistical Society, A, 159*, 149–63.

Gray, J. Jesson, D., Goldstein, H., Hedger, K., & Rasbash, J. (1995). A multilevel analysis of school improvement: Changes in schools' performance over time. *School Effectiveness and School Improvement, 6*, 97–114.

Hamilton, D. (1996). Peddling feel-good fictions. *Forum, 38*, 54–56.

Hill, P.W., & Goldstein, H. (in press). Multilevel modelling of educational data with cross classification and missing identification for units. *Journal of Educational and Behavioural Statistics*.

Hill, P.W., & Rowe, K. J. (1996) Multilevel modelling in school effectiveness research. *School Effectiveness and School Improvement, 7*, 1–34.

Jencks, C. S., Smith, M., Ackland, H., Bane, M. J., Cohen, D., Gintis, H., Heyns B., & Micholson S. (1972). *Inequality: A reassessment of the effect of family and schooling in America.* New York: Basic Books.

Longford, N. T. (1987). A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested effects. *Biometrika, 74,* 812–27.

Marks, J., Cox C., & Pomian-Srzednicki, M. (1983). *Standards in English schools.* London: National Council for Educational Standards.

Mortimore, P., Sammons, P., Stoll, L., Lewis, D., & Ecob, R. (1988). *School matters.* Wells: Open Books.

Myers, K. (1995). *School improvement in practice.* London: Falmer.

Myers, K., & Goldstein, H. (1996). Get it in context? *Education, 16 February,* 12.

Rasbash, J., & Woodhouse, G. (1995). *MLn Command reference.* London: Institute of Education.

Raudenbush, S. W., & Willms, J. D. (1995). The estimation of school effects. *Journal of Educational and Behavioural Statistics, 20,* 307–336.

Riley, K. A., & Nuttall, D. L. (Eds). (1994). *Measuring quality.* London: Falmer Press.

Robinson, W. S. (1950). Ecological correlations and the behaviour of individuals. *American Sociological Review, 15,* 351–7.

Rowe, K.J., Hill, P.W., & Holmes-Smith, P. (1995). Methodological issues in educational performance and school effectiveness research: a discussion with worked examples. *Australian Journal of Education, 39,* 217–48.

Rutter, M., Maughan, B., Mortimore, P., Ouston, J., & Smith, A. (1979). *Fifteen thousand hours.* Wells: Open Books.

Sammons, P., Mortimore, P., & Hillman, J. (1996). Key characteristics of effective schools: a response to "Peddling feel-good fictions". *Forum, 38,* 88–90.

Sammons, P., & Reynolds, D. (1996). A partisan evaluation: John Elliot on school effectiveness. *Cambridge Journal of Education, 26,* to appear.

Sanders, W. L., & Horn, S.P. (1994). The Tennessee value-added assessment system (TVAAS): mixed model methodology in educational assessment. *Journal of Personnel Evaluation in Education,* 8, 299–311.

Scheerens, J. (1992). *Effective schooling: Research, theory and practice.* London: Cassell.

Woodhouse, G. (1996). *Multilevel modelling applications: A guide for users of MLn.* London: Institute of Education.

Woodhouse, G., & Goldstein, H. (1989). Educational performance indicators and LEA league tables. *Oxford Review of Education,* 14, 301–319.

Woodhouse, G., Yang, M., Goldstein, H., Rasbash, J., & Pan, H. (1996). Adjusting for measurement unreliability in multilevel analysis. *Journal of the Royal Statistical Society, A, 159,* 201–12.