Title	Evaluation of school performance and the school effectiveness debate
Author(s)	Sim Yoke Hwee and Poh Sui Hoi
Source	*ERA - AARE Joint Conference, Singapore, 25-29 November 1996*

EVALUATION OF SCHOOL PERFORMANCE AND THE SCHOOL EFFECTIVENESS DEBATE

by

Ms Sim Yoke Hwee andDr Poh Sui Hoi
Administrative OfficerSenior Lecturer
Educational & Staff Development Dept.Psychological Studies Division
Singapore PolytechnicSchool of Education, NTU/NIE
500 Dover Road469 Bukit Timah Campus
Singapore 139651Singapore 259756
Email: S30177@SP.AC.SG Email:POHSH@NIEVAX.NIE.AC.SG

Paper given to ERA-AARE Joint Conference
on
Educational Research: Building New Partnerships

25 - 29 November 1996, Singapore Polytechnic, Singapore

Abstract:

This exploratory study attempts to call into question the effects of schools on educational attainment, especially at ages 16-19 years of age. The principal questions investigated were:
To what extent do schools vary in their A-level performance?
Are schools differentially effective for pupils of differing abilities?
What was the significance of contextual variables in accounting for school effects on pupil achievement?

Data on 2541 pupils from 33 schools were analysed by the recent multilevel statistical technique. On methodological grounds, the study was distinguished as part of the third wave movement of school effectiveness.

Results showed that some of the schools were indeed more effective educational environment than others. Of the total variance in the 'model', an estimated 6% were

accounted between schools.

Acknowledgements:

The authors would like to thank Mr David Jesson, Director of the Performance Indicators Project
at the University of Sheffield, UK for his advice and wise counsel. We also would like to thank our sponsors to this conference, the Singapore Polytechnic and the Nanyang Technological University/National Institute of Education, Singapore.

INTRODUCTION

In recent years, the league table debate has brought out diverse initiatives for assessing schooling quality. Different educational circles have different evaluative criteria; but rightly or wrongly, examination results are commonly perceived as the traditional yardstick of educational success (Gray,1990). Legislative requirement to publish exam statistics has forced the issue of performance indicators back on the boil with a renewed cachet in research terms and attempts to integrate them within the service are now urgent and assured.

Knowledge gleaned by the School Effectiveness literature has something to contribute here. There are currently two basic frameworks in use for interpreting schools' exam results (Jesson, 1992). The easiest way is to compare schools' raw scores with each other or with either a local or national average (the Standards Model) to make statements about the effectiveness of each school. Given that pupils are not randomly assigned to 'treatments' (or schools), the Standards Model runs the distinct risk of rewarding schools for eliciting good results from the quality of the intakes they (can) attract rather than what they actually do with their pupils. In contrast, the Progress Model and its variants present exam results in the context of the kinds of students attending each institution and, hence, allow like-with-like comparisons. This latter 'value-added' approach is widely advocated by researchers in the field.

OVERVIEW OF THE PROBLEM

Existing Government policy to publish only raw results in standard tables is likely to endanger the quality of educational provision because schools' levels of effectiveness are not measured properly. Information on pupils' absolute attainment, such as the percentage of A-levels achieved at grades A and B, is informative but conveys little about the contribution institutions make to pupil progress. Such statistics are not a good guide to help identify schools in dire need

of improvement and may even breed complacency amongst some. Schools endowed with bright, motivated young people on intake can cover up for performance that may be below expectation through "good" exam results whilst those with "poor" grades may justify their low grades simply because they serve socially and economically disadvantaged pupil populations. Only by assessing the value added by a school, using appropriate statistical analytic procedures, can fair comparisons between them be made and subsequently used to inform parents, professionals and others about choice.

Ideally, information about prior achievement on entry to a school set against information of attainment on exit should provide a measure of the progress pupils make during their time at that school. Other input variables are also known to be significant and need to be taken into account. For example, boys perform better than girls at A-level (Fitz-Gibbon, 1989; Gray and Jesson, 1991), so it would be inappropriate to compare a single-sex school with a mixed school without considering the gender effect.

However, several difficulties that beset some previous research on School Effectiveness can readily be identified. These studies have been criticised on the grounds of the paucity of their intake measures and because they collected and used data at the level of the school, rather than the individual student. This entails aggregation errors and threatens one of the key assumptions on which such studies are premised, namely, that major differences in the intakes to different schools have been taken into account in the analysis. Hence, unless

comprehensive background pupil data are available, research in the field is unlikely to be able to adequately compare like-with-like.

Secondly, prior to Aitkin and Longford's seminal work in 1986, statistical procedures used to treat data were predominantly based on multiple regression analyses involving single-level designs. These traditional techniques did not reflect the logically hierarchical nature of the variables that describe students (micro level) nested within schools (macro level). Recent statistical advances in multilevel methods (e.g., Aitkin & Longford,1986; Goldstein, 1986) fit the structure of school data better, making it possible for both student-level and school-level data to be analysed together in a multistage model.

Furthermore, by using a two/multi-level investigation, the school effects can be separated from the student effects in accounting for variance in pupil achievement. This greater efficiency in the use of indicator data allows more precise and unbiased estimates to be made of the potential influence of schools on pupils' learning outcomes.

PURPOSE OF THE STUDY

The exploratory study undertaken here attempts to call into question again, the effects of schools on educational attainment; specifically in the post-compulsory phase at 16-19 years of age. The principal questions to be addressed are:

(i) To what extent do schools vary in their A-level performance?
(ii) Are schools differentially effective for pupils of differing abilities? and,
(iii) What is the significance of contextual variables in accounting for school effects on pupil achievement?

Multilevel modelling was used as the method of analysis. Several variables describing schools and students were incorporated in the procedure at two levels. These input variables represent attributes over which schools had little or no control but were expected to have influence on the outcome measure. Various different analyses were carried out in an exploratory manner before arriving at the most efficient set of input and outcome variables for inclusion in the final model.

The discrepancies between the predicted and observed outcomes are known as the residuals or "school effects". These estimates may represent in part the effects of school policies and practices largely under the control of the schools and not included explicitly in the regression equations. Thus, a school whose presumed effect on the outcome is average would have a residual near zero. In the same vein, a school with a strong positive effect on the outcome is associated with a relatively large positive residual and vice versa for another whose effects on the outcome falls below that of the predicted value would have negative residuals. But, since all prediction models incorporate 'errors' in estimation, these also need to be taken into account when estimating a school's effectiveness.

A further objective of the study is to compute the different sets of residuals arising from the null model and the variance components model in its variant forms. The procedures organised schools into rank orders based on their effect sizes in a value-added paradigm. In the final stage, these results were compared with those obtained from a raw-score ranking of the same schools in the study.

REVIEW OF RELATED LITERATURE

In reviewing the research literature on the effects of schools upon their pupils, several significant 'high points' in its historical

development were observed. These are often the manifestation of some statistical or methodological advances, or the accumulation of some significant findings which had ushered fellow workers within the field into a 'paradigm shift'. Upsurge of studies in the new paradigm followed until the next 'crest' came along. These growth patterns have, therefore, being termed 'waves', using the 'wave' analogy popularised by Toffler (1981). The researchers have now entered the 'third wave' in the development of the movement with the advent of technical and conceptual advance offered by hierarchical linear models.

THE FIRST WAVE: 'SCHOOLS MAKE LITTLE DIFFERENCE'

The much publicised Coleman Report (Coleman et al., 1966) should be seen as one impressive starting point of the school effectiveness tradition, although the main thrust of the project was to investigate (in)equality in education. Extensive data collected from 4,000 schools showed that there were substantial differences between the verbal ability levels of pupils from various minority races and those of white children. After statistical adjustments had been made for the influence of home background, the researchers found that schools accounted for only an approximate 10% of the variance in pupil achievement. This led to the assertion that the apparent effects of schools upon pupil attainment were merely a reflection of the social composition of the schools' pupil bodies and not a reflection of the quality of the schools themselves.

The Coleman data together with other longitudinal data were re-analysed by Jencks and his associates. In most respects, the results of this second major study (Jencks et al., 1972) parallel those of the first: that a much higher proportion of variance between individuals' ability scores could be explained by their socio-economic status and level of intelligence rather than to measures associated with material quality of the schools. Set alongside the weight of similar evidence accumulated in Britain by the Plowden Committee (Department of Education and Science, 1967), the combined effect was a substantial erosion both in public confidence in education and educators and in educators' belief in their own efficacy.

Detailed criticisms of these early studies, both in terms of methodology and of substantive findings began to emerge from the late sixties onwards (for instance, Dyer, 1968; Madaus et al., 1980). Amongst other criticisms, doubt was also cast on the analysis technique of multiple regression predominantly used. Typically, a series of sets of variables (e.g., family or school measures) were entered into the analysis in logical sequence, thereby permitting an estimate of the total variance that was due to the various predictors. Such a procedure carries a potential source of error: if there is a high inter-correlation between the (family and school) measures, the predictive power assigned to background factors is that which they

possess uniquely as well as the explanatory power they share with school resources (Smith, 1972).

In view of these shortcomings, some researchers performed re-analysis work on the databases of the studies themselves. Such assessments threw

up new insights: amongst others, the schools in the Coleman Report had a quite marked effect upon children as they grew into adolescence; and the school factors were shown to be highly inter-correlated with certain non-cognitive outcomes of the educational process (Dyer, 1968). Developing alongside these secondary analyses of data was the emergence of a growing body of literature which purported to show that - contrary to popular belief in the tide of pessimism - some schools were actually more effective as educational environments than others.

THE SECOND WAVE: 'SCHOOLS MAKE A DIFFERENCE'

One of the most influential work that succeeded in informing the view that the effects of schools are considerable (even when account is taken of differences in intake), was that of Rutter and his team (Rutter et al., 1979). The longitudinal study carried out from 1970 to 1974 was welcomed as a boost to the battered morale of the teaching profession since it sought to show that it was within the practitioners' grasp to make schools more effective. Like its predecessors, the study was not spared from vigorous criticism (e.g., Acton, 1980; Goldstein, 1980). Whether or not all the potential intake effects on schools had been controlled for was the main line of attack.

The technical limitation at this stage of development of the research paradigm had also meant that many studies in the 'second wave' had resorted to higher level analyses using schools or LEAs as units instead of individual-based pupil statistics. Thus, the Inner London Education Authority (ILEA, 1980) carried out a longitudinal study of average school examination results using average school intake test scores. Such designs, though not entirely useless, are limited in their insights and knowledge offer, as pointed out by Goldstein (1984) in an article arguing for a full multistage study design:

> ... relationships at, say, the individual level may be quite different from the relationships which exist between the same variables measured at the school level. Thus there may be differences in rates of progress between different types of school using pupils but not using schools as units. In the absence of suitable data at all levels, therefore, we are not able to make, in general, useful inferences about lower level relationships from only higher level data. (p73)

THE THIRD WAVE: MULTILEVEL MODELLING METHODS

The traditional analytic method used hitherto in research studies of the First and Second Waves requires the assumption that educational interventions have a constant effect on all students who are exposed to them, and that these effects are invariant across organisational contexts. In reality, subjects are 'nested within layers' (Barr & Dreeben, 1983) of classrooms, schools, education authorities and local communities so that responses within clusters and across contexts are dependent. An attempt to portray this 'nestedness' of educational data is shown in Figure 1.

Figure 1, about here

A number of methodologists, working independently, have developed estimation procedures appropriate for the hierarchical nature of social data. Seminal work in this area by Aitkin and Longford (1986),

Goldstein (1986), de Leeuw and Kreft (1986), and Raudenbush and Bryk (1986) have ushered in the application of multilevel modelling techniques into school effectiveness studies. Although an extension of ordinary multiple regression methods, the technique is superior over its predecessor for at least two main reasons, one substantive and the other technical (Paterson, 1989). The substantive one is that a multilevel regression enables the formulation of explicit modelling of processes occuring within and between educational units by disaggregating the relationship between the 'response' and the 'explanatory' variables into separate parts (two for a two stage model, etc). Hence, for instance, the within-school component is addressed by comparing the performance of individuals who attended the same school while the between-school element takes account of differences among schools. In a final stage of synthesis, these differences are explored in terms of school characteristics.

Allowing individual relationships to vary among groups is necessary for technical reasons as well because grouping may introduce an extra random component and thus have an effect on the standard errors. Multilevel methods permit the incorporation of such group characteristics into models of individual behaviour, while also yielding statistical tests of whether the groups in the sample are really different from each other, as well as confidence intervals for the extent of any differences.

The quarrel over the appropriate 'unit of analysis' (Burstein, 1980) was also to be amiably settled by the multilevel models which have been proposed under a variety of names: variance component model, mixed models and hierarchical linear models. Much of the credit for the

advance on this particular statistical modelling issue must be given to Aitkin and Longford (1986). In what can be described as an excellent exposition of the contrasting results obtained by fitting different models to hierarchical data, the statistical robustness of the variance component method (described as their Model 5) was explicitly demonstrated.

Nuttall et al. (1989) in using the multilevel modelling method to analyse exam results at 16+, have explored the issue that schools differ in terms of the achievement of students from different ethnic minorities, and that some schools have the effect of compensating for initial differences whilst other schools have the effect of magnifying them. Jesson and Gray (1991) in a more recent work have used such models to further investigate this issue of differential effectiveness: their various analyses led them towards the conclusion that even though there was some evidence of differential effectiveness, it was of a distinctly limited nature; the position whereby schools which were more effective for a sub-group of pupils were also more effective for other groups represents the 'centre of gravity'.

This brings the debate to the present-day where the atmosphere of testing and reporting school results has led to a frenzy of performance indicators and the interpretation of the published statistics as they stand. The contention is between evaluating school performance according to reports of pupils' absolute levels of attainment and the value added framework which measures the progress students have made whilst in school. Although Government league tables have persisted in their 'crude form', the principle that examination results should be subjected to value added analysis is slowly but surely gaining currency.

As an endpiece to this literature review, salient characteristics of the three waves of school effectiveness research and debate are

presented in a summarised form shown in Figure 2.

Figure 2 , about here

METHODOLOGY

The study uses an existent dataset supplied by courtesy of Mr David Jesson, Director of the Performance Indicators Project at the University of Sheffield (1991). As guarantees of confidentiality are absolutely essential in research of this nature, all data are entered in numeric form, hence, the units studied (whether pupils or schools) are distinguishable only by their numeric codes. The data reside in an

ML3-E worksheet containing 17 field entries. Basically, the database contains the past General Certification School Examinations (GCSE) and present A-level results of 2541 students who have participated in ages 16-19 post-compulsory education. The pupils are drawn from 33 institutions covering several major categories of A-level provision: Local Educational Authority (LEA)-maintained and Grant-maintained secondary schools with sixth forms, Sixth-form colleges and Further Education (FE) colleges. The examination candidature varies considerably across institutions. The overall gender ratio is 4 boys (coded 1) to 5 girls (coded 0).

In order to explore outcome differences between institutions with a view to describing their relative effectiveness, detailed information about relevant student characteristics as well as their initial and final attainments were taken into consideration. To better reflect the questions addressed by the study, further data description was taken up via two sub-categories:
measures of student intakes, and
(ii) measures of educational outcomes.

MEASURES OF STUDENT INTAKES

The two intake variables available for creating a 'level playing field' on which to base school comparisons are pupils' gender and prior attainment at GCSE. Both of these factors have been conclusively shown by previous studies to be statistically important for prediction of pupils' A-level results (Fitz-Gibbon, 1989; Gray & Jesson, 1991) although it must be mentioned that the influence expressed through initial achievement at GCSE (or its preceding equivalents) far exceeds that of gender.

Pupils' intellectual ability at entry to sixth-form is given in various constructions: GCSEPTS (total grade scores), GCSEAVGE (GCSE average score per subject) and BEST5 (sum of grade scores in English, Mathematics and the best 5 other grades or all other grades, if fewer than 5 other subjects was attempted). The reported scores were scaled by converting the results of each subject using the formula:

GCSE Grade	ABC	D	E	F	G	U
Score	76	5	4	3	2	1	0

To assess the predictive power of these measures, a series of regressions was carried out of A-level results (ALEVPTS) against data in each column in turn. Though GCSEAVGE yielded the highest correlation

coefficient (0.6730), it is not favoured as an explanatory variable for

analysis because the averaging of score attainment in all subjects loses information and 'collapses' the actual variability in pupils' initial starting points. For the remaining two variables, BEST5 emerged with a higher correlation coefficient of 0.5886 (Table 1). For these reasons, and following the substantive ground not to put too high a premium on the number of subjects taken at 16+, the BEST5 measure is a preferred summary of pre-existing differences between school intakes.

Table 1, about here

An additional point is that only one indicator of prior ability needs to be considered in the regressions because the given measures are not mutually exclusive of each other but are highly correlated. The descriptive statistics in Table 1 show that for pupils who stay on for A-level studies, their average BEST5 GCSE score was approximately 38 points with a standard deviation of 6.28. With a staggering 90% of candidates scoring 5 B's or better, some sort of entry selection process to sixth-form studies is apparently in operation. This observation is in full accord with some findings of the Youth Cohort Study which has explored the motivations youngsters cite for staying on in full-time education (Jesson et al., 1991; Gray et al., 1993). Among other considerations,

> ... the formal qualifications young people obtained in the examinations they sit at the end of their period of compulsory schooling are overwhelmingly the most powerful predictors of further educational participation. In general, we have found the relationships to be linear ones. The better their qualifications, the more likely a young person is to stay on. (Gray, Jesson & Tranmer, 1993, p4)

Besides pupil-level variables, school-level variables such as school size or school means on intake scores or other group characteristics derived from the aggregation of pupil data within each school are also eligible candidates for explanatory variables. The latter is an example of the 'context' or 'compositional' variables which, if found significant, might offer evidence about the so-called 'balance thesis' (Willms, 1985) which argues for the effect of school context on the achievement of an individual pupil. Table 2 shows the distribution of this factor (for BEST5) which has a school-level mean of around 37 points and a standard deviation of around 3. This possibility was explored in the study by including in the model specification, the school means on BEST5. Some caution must be exercised, however, in the interpretation of such 'compositional' variables, since although they may be associated with more 'powerful' models at school level, the

factors themselves may appear more important than they actually are by reason of incomplete, misspecified or poorly estimated factors at pupil level.

Table 2, about here

MEASURES OF EDUCATIONAL OUTCOMES

The number of Advanced-level and Advanced Supplementary-level subjects students were entered for are summarised by the histograms in Table 3 below:

Table 3, about here

45% of pupils sat for the modal number of 3 A-level subjects whilst interestingly, some 2% of pupils who participated in the courses did not sit for exams. Participation rate at AS subjects was a mere 14% of the total candidature; of this, 78% signed up for only one of such examinations.

For scoring A-level grades, the UCCA scheme is used:

A-level Grade	A	B	C	D	E	N	U
Score	10	8	6	4	2	0	0

Note: Grade N denotes narrow failure
Grade U (unclassified) replaces grade F

On performance in A-level papers, a pupil was most likely to obtain at least 2 Cs or an average point score of 12.8. A substantial proportion of students (42%) who sat for the supplementary-level papers scored modestly between 1 to 3 points inclusive. Results in specific curriculum areas are not supplied except for that of General Studies which can readily be worked out from data in three columns: ALEVPTS, ASPOINTS and A&AS-GST (GST is taken by at least 26% of the pupils and had a mean of approximately 5.3).

Only cognitive outcomes in the public examination are considered for analysis. This is not to deny the importance of 'social' or other non-cognitive outcomes of education such as students' attendance or attitudes but simply so that attention may be focussed on one of the primary outputs of education, namely, pupils' examination achievements.

Obtaining academic success takes on greater significance in study at sixth-form because almost all students who stay on this post-compulsory phase do so to secure qualifications.

As with the prior attainment variables, a choice of several outcome variables is given. They are ALEVPTS (sum of scores for A-level subjects), MEANPTS (the average score per A-level subject), ASPOINTS (total score for AS-level grades), MEANASPT (the mean score per AS-level subject) and A&AS-GST (overall score for all A and AS-level subjects excluding results in General Studies). The latter column is given because some analysts have commented that results in General Studies should not be included for comparative evaluation of institutions' courses: (i) the results do not reflect attainment reached by following specific courses of preparation similar to those for particular disciplines, and (ii) not all students who are capable of passing these exams attempt them (Audit Commission & OFSTED, 1993). The exclusion does not call into question the value of the particular subject option.

From ALEVPTS and ASPOINTS, a different outcome variable which embrace both A and AS level points is easily constructed and set into one of the columns (named A&ASPTS). All six field entries can potentially be used as response variables, but only ALEVPTS, A&AS-GST and A&ASPTS offer a more comprehensive coverage of the subjects attempted. Hence,

subsequent analysis and explorations will be restricted to data of just these three 'shortlisted' predictors.

In Table 4, some further summary statistics for each school is presented; namely the mean and standard deviation for both the BEST5 and A&ASPTS score. A unique feature of the tabulated figures is the relative homogeneity of the sample, i.e., schools are broadly comparable in their intake ability (except for slightly lower prior scores in schools with ID codes 29-33).

Table 4, about here

Most schools (82%) are within one standard deviation of the BEST5 average taken across institutions. By this criterion, most of the institutions are neither particularly advantaged or disadvantaged by the quality of their intakes.

RESULTS

The intake measures which are to be considered for analysis are GENDER

and BEST5 at pupil level and mean BEST5 at school level, whilst three outcome variables (ALEVPTS, A&ASPTS and A&AS-GST) have been shortlisted. A decision must at this stage be made on the choice of response variable since, firstly, all three are related cognitive outcomes; and secondly, only one of such measures can be modelled at a time. A series of estimation routines was conducted for each of the potential candidates, keeping BEST5 and GENDER as the explanatory variables in each run. From estimates in the random part, the explanatory power of each model, expressed in terms of R2 is also obtained. Table 5 summarises the results.

Table 5, about here

All things being equal, the model with A&ASPTS as response variable (and BEST5 and GENDER as explanatory variables) explained the greatest amount of total variance in A-level performance. On this ground, A&ASPTS becomes the default choice of outcome variable for the study (it is noted that an identical structure exists for the pupil-level variable GCSEPTS). Substantive reasons can also be cited to support it: the comprehensive coverage of take-up in all A and AS subjects is responsive to concern from many quarters, including the government and employers, to promote greater breadth in A-level study (Department of Education and Science, 1987).

VARIANCE COMPONENTS MODEL

Various models were applied to the set of data. Computation progressed logically from the 'base case' or 'null' model to modelling with fixed effects and subsequently, extension to allow for random slopes. In the light of these analyses, residuals distilling from each model are used to evaluate the relative effectiveness of schools in the sample.

THE NULL MODEL

Beginning with the simple two-level 'base case' in which no explanatory variables are included, the variance components estimates are:

School-level Variance = 9.16
Pupil-level Variance = 93.79
Total Variance = 102.95

Thus, about 9% of the total 'raw' variance is between schools and 91% among pupils within schools. These figures indicate that the schools in

the dataset were a fairly homogeneous sample (see Table 6)

Table 6, about here

MODELLING WITH FIXED EFFECTS

In the second run of the analysis, pupil-level variables (namely, GENDER and BEST5) were fitted. In addition to modelling fixed effects (which are constant across schools), the multilevel program investigates the presence and nature of random effects (which vary from school to school). In the random part, the previous estimates of variance at each level were markedly reduced and are now:

School-level Variance = 2.312
Pupil-level Variance = 62.42
Total Variance = 64.732

Hence, nearly 75% {(9.16-2.312)/9.16} of the school variance component, but just over 33% {(93.79-62.42)/93.79} of the pupil variance component, is attributable to the explanatory variables. Expressed differently, R2 for the model is 37% {(102.95-64.732)/102.95} for total variance, 33% for pupil variance and 75% for school variance. The smallness of the proportion at school-level of variance (less than 4%) is at the lower end of estimates from previous work and is very much in line with the more conservative range of estimates (albeit for different school situations) assessed by recent multilevel frameworks (Aitkin & Longford, 1986; Gray et al., 1986; Jesson & Gray, 1991). Both the pupil-level variables showed significant coefficients but the predictive power of the former overshadows the gender effect (see Table 7).

Table 7, about here

Level-1 variables are normally used in their linear form but transformations of these variables can also be entered (see Goldstein, 1987, p27). Indeed, for this dataset, when a quadratic functional form of the prior achievement measure (BEST52 labelled BESTSQ) was added to the existing model, a substantial amount of the remaining pupil-level variation was explained. The additional variable was found to be statistically significant.

The squaring of BEST5 makes substantive sense in that the A-level examination is designed for the more able students and one would expect the attainment of high grades at 16+ to be particularly relevant to

subsequent performance at A-level. Squaring the variable enhances the effect of the higher (A, B or C) grades at GCSE.

The third step was to fit the school-level variable, the mean of BEST5 (labelled AVBEST5). It is interesting to note that with this transformation of pupils 'ability' scores, the contribution made by the compositional variable was sharply reduced. This proxy for school context did not fall within the 5% significance level. There was, therefore, less evidence for a strong positive relationship between examination performance and school composition. The contextual variable was, henceforth, removed from subsequent analyses.

MODELLING WITH EXTENSION TO RANDOM SLOPES

The intercept variance (CONSTANT/CONSTANT) is estimated to be 3.42. This figure is more than twice its associated standard error, suggesting again that schools really do have different intercepts. The correlation of the slopes and intercepts is estimated to be 0.727 {(0.176/(3.42*0.0171)0.5}, so it appears that students with higher prior ability have a moderately high tendency to perform better in A-level study. The most interesting finding is, however, that the schools' coefficient of BEST5 (0.0171) was just significant at the 5% level. There is, therefore, evidence for differential slopes existing within the schools of this dataset. The breakdown of this evidence by schools shows clear statistical evidence in favour of it in 5 institutions (Schools 1, 3, 10, 17 and 29). Table 8 shows the differential slopes.

Table 8, about here

The slope configuration of these five cases is one whose curves 'fan out' from each other, similar to that observed in Jesson and Gray (1991). (This configuration is expected due to the positive correlations of slopes and intercepts). The extent of divergence from the overall curve gives an indication of the degree of differential effectiveness occurring in any school. A visual documentation of the phenomena is presented in Figure 3.

Figure 3, about here

In general, however, most of the 'performance' curves do not diverge and would have been parallel if there were no quadratic terms in the

specified model, showing that the majority of schools in the sample have similar effects on most of their pupils. To illustrate the issue further, the regression curves of two schools have been graphed (See Figure 4). These schools were chosen on the basis that both have significant slope and intercept differences arising from the analyses.

Figure 4, about here

Hence, their curves not only diverge from the overall case but are also physically displaced from it. The respective curves for the hypothetical position where the schools' performance were not differential are also included (indicated by 1b and 3b). These curves crossed the previous ones (indicated by 1a and 3a), suggesting that for School 1, it is more effective for the abler pupils but it may

disadvantage the less able ones.

Finally, the model of pupil performance at A-level can be explicitly presented:

Examination Score = 10.04
+ 1.238 x (BEST5-38)
+ 1.721 (if pupil is male)
+ 0.043 x (BEST5-38)2

and R2, the explanatory power of the model is calculated to be 50.1% {(102.95-51.36)x100/102.95}.

COMPARISON OF RANKING ORDERS

With variation in intake allowed for, the models can reasonably be used to make comparisons between institutions in their performance. Rank orders can be established based on the school-level residuals; these really are the discrepancies between schools' observed and predicted outcome scores. A fine ordering has been presented (See Table 9) for the most efficient 'fixed effects' model which is specified by GENDER, BEST5 and BESTSQ as explanatory variables. This ranking of intercept residuals is preceded by School 1 as the best 'performer' while School 3 did worst with its pupils. The average score difference between them is over 7 points or at least an additional C grade at Advanced-level for the median pupil in School 1 compared with the same pupil's predicted performance in School 3.

Table 9, about here

In the same context, the estimated range in the 'value added' score that could be typically expected of an average pupil in each school was computed. This range was calculated by employing the 95% confidence interval for each school's score. In Figure 5 below, this statistical measure is indicated by vertical bars while the horizontal ticks represent the average raw scores in A&ASPTS for each school.

Figure 5, about here

For more than half (19 out of 33) the sampled schools, the actual results were within the confidence intervals of their predicted scores. These institutions were really performing as might be expected for their mix of pupils. Conversely in the rest of the cases, some schools did significantly better while some did significantly worse than their predicted level. This indicates that one cannot safely or confidently differentiate between the performance of these schools (Goldstein & Healy, 1993); they are doing as well as expected. As a rule of thumb, only those schools with ranges which are different from zero have clearly different results, and so, a detailed school ranking like those proposed in Table 9 does not make statistical sense. This is actually true regardless of whether raw results or value-added results are adopted. In practice, therefore, preference should be given to a categorical positioning of schools where schools of broadly similar results are not distinguished as performing better or worse. In this study, if the sampled institutions are divided into three sub-groups, say, 'A' as equivalent to the top 25% (or 8) schools, 'B' for the middle 50% (or 17) schools and 'C' for the lowest 25% (or 8) schools,

most (29 out of 33) schools do not change categories if the null model is used instead of raw results, as shown in Table 10 below.

Table 10, about here

In stark contrast, if account is taken of initial pupil variability as in the 'fixed effects' and 'random effects' models, 42% (14 out of 33) of schools changed categories from A to B or from B to C. In other words, if one insists on file-ordering schools by their crude scores, the probability of appropriately estimating a school's relative performance is very poor indeed, and what is more, half the schools indicated by the raw scores as in the 'top' 25%, are in reality performing only at average levels, whilst similar observations apply to those ranked 'lowest' by their raw scores. Even those ranked conservatively, as in the middle 50% are wrongly allocated almost as

often as that are correct.

DISCUSSION AND CONCLUSION

- The Impact of Schools on Pupil Performance

In assessing the effect of various input variables explicitly included in the regression equations, GENDER, BEST5 and BESTSQ were statistically significant. The impact of prior achievement on the outcome is classical and well established. The same can be said for the gender effect at this stage of schooling. Starting with similar exam achievements at 16+, male students do marginally better than their female counterparts at A-level. The three variables, in combination, explained a substantial proportion of the total raw variance in the null model.

Nonetheless, there is still an element of pupil performance which is attributable to schools after intake adjustments have been made. For the model defined by GENDER, BEST5 and BESTSQ, the variance between schools was estimated at around 6% and was highly significant. In order that the separate impact of individual schools on student achievement can be established, membership of pupils in the school of A-level study was included in the analyses. Those schools with the largest positive and negative intercept residuals were identified as performing better or worse than would be expected given the characteristics of their pupils. The average score difference between them was at least 7 points or an additional C grade at A level for a median pupil in the most effective school compared to the same pupil's predicted performance in the least effective school.

- On Differential Effectiveness

The results of analyses of the effects of individual schools on pupil exam outcomes indicated that in general, there were few differences between students of different abilities in these effects. To put it differently, schools which were effective in promoting achievement for a particular ability group tended also to be effective across-the-board and vice versa. It does appear, however, that in this sample, clear statistical evidence for differential effectiveness existed within a small number of the schools. Among these schools, three did better in assisting their abler pupils to achieve exam success whilst appearing to depress the performance of less able ones. The reverse situation was true of the other two educational environment.

- On Contextual Effects

When the proxy for school context was included in the equations, its

contribution was almost significant at 5% level. Some caution must be exercised, however, in the interpretation of such compositional variables: the apparent effect could reflect remaining variation at Level 1 rather than a true Level 2 effect. This conjecture would be supported if the Level 2 effect disappeared when additional Level 1 variables, or transformations of existing Level 1 variables were included in the model. When the contextual variable was added to the set of explanatory variables, it showed a significant positive effect. The variance explained increased slightly from 37.1% to 37.3%, while the variance remaining between schools was reduced from 25% to 23%. When BESTSQ was fitted, the variance explained by the model increased further from 37.3% to 49.6%, but the contribution made by the contextual variable was sharply reduced so that it no longer featured as a significant factor in this model.

- 'Value Added' versus 'League Table' Evaluations

The findings in this study uphold the value added approach as the more appropriate form of comparative evaluation. Nevertheless, even with value added analyses, the interpretation of school differences needs to be approached carefully. The estimates of the unknown school effects are a compound of errors in specification, in estimation as well as incorporating genuine differences between schools' performances. It is also the case that many of these estimates tend to have relatively large standard error estimates (particularly for schools with very small numbers of students) so that the performance of these schools cannot be separated reliably from each other. This situation was succinctly illustrated by the substantial overlap of confidence intervals of the residuals for many of the schools in the sample. Carrying the concept further, even though a difference may exist numerically in calculations, no substantive meaning can be attached to that difference if it is "too small". On this ground, a fine rank ordering of schools as is often found in crude league tables, is inappropriate and does not make statistical sense. A more conservative view of these differences is taken if one categorised schools into only broad performance bands, as suggested earlier. Statements of effectiveness or otherwise of schools can at best be made for extreme schools which may be performing convincingly better or worse than their predicted level.

- Concluding Comments

At the close of the study, the following conclusions need to be reiterated:
The school to which a student belongs has a significant effect on that student's exam outcome. Of the overall variance in A-level scores, an approximate 6% lies between schools after account has been taken of major explanatory variables;

Statistical evidence in favour of differential effectiveness was

detected for several schools within the sample population. In general, however, most schools were equally effective or ineffective for the whole ability range of their pupils;

The contextual variable was not significant. Thus, the composition of a school's intake did not have substantial influence on pupil performance over and above the effects associated with an individual's ability and gender.

In comparing the assessments of school effects made by 'value added' and raw results, it is clear that raw results were inappropriate ae indicators of schools' relative performance. Rank orders based on the two sets of statistical information were alarmingly disparate, pointing to the urgency to discount the flawed league table approach.

The study was almost exclusively based on quantitative data supplied from an external source. As such, it was appropriate for the limited purposes addressed. It does, however, have a number of limitations. The most important of these relates to its omission to reference to the field of 'school improvement'. This area has its own, rather distinct, disciplines and methods and relies to a considerable extent on micro-studies of the processes and practices of particular institutions - particularly those undergoing change or where there is thought to be evidence of either 'good' or 'bad' practice.

It would also have been interesting to have had access to subject area data to assess the extent to which schools had 'strengths' and 'weaknesses', that is, whether their effectiveness was an 'across the board' phenomenon or whether different subject areas were very different in their effectiveness.

The last issue concerns the measure of effectiveness used in this study which is essentially a 'steady state' or 'one-off' account, using only one year's data. Clearly, schools will vary in their performance over time; schools which retain effect over two or more years are obviously more effective than those which appear to be so in one year but not in the next. It would be helpful, therefore, to build further cohorts of data into an expanded database to explore such issues in the context of performance at A level.

Finally, notwithstanding these limitations, the model of pupil performance that has been specified here was adequate in that it explained a substantial portion of the total variance in schools' outcome scores. That being so, the study has served its intended purpose and has pointed the way to possibilities for future research in this important area.

REFERENCES AND BIBLIOGRAPHY

Acton, T. A. (1980) 'Educational criteria of success', Educational Research, 22, 3, 163-169.

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

Audit Commission & OFSTED (1993) Unfinished Business: Full-time Educational Courses for 16-19 Year Olds, London: HMSO.

Barr, R. & Dreeben, R. (1983) How Schools Work, Chicago: University of Chicago Press.

Burstein, L. (1980) 'The analysis of multi-level data in educational research and evaluation', in Berliner, D. C. (ed.) Review of Research in Education, Washington DC: American Educational Research Association, 8, 158-233.

Coleman, J. S., Campbell, E. Q., Hobson, C. Y., McPartland, Y., Mood, F. D., Weinfeld, F. D. & York, R. L. (1966) Equality of Educational Opportunity, Washington, DC: US Government Printing Office.

de Leeuw, J., & Kreft, I. (1986) 'Random coefficient models for multi-level analysis, Journal of Educational Statistics, 11, 1, 57-85.

Department of Education and Science (1967) Children and their Primary Schools, London: HMSO. (The Plowden Report)

Department of Education and Science (1987) Examination Reform for Schools: A Guide for Employers to Recent Changes in the School Examination and Assessment System, London: HMSO.

Dyer, H. (1968) 'School factors and equality of educational opportunity', Harvard Educational Review, 38, 38-56.

Fitz-Gibbon, C. T. (August 1989) Multilevel Modelling in an Indicator System Paper presented to the ESRC International Conference on Multilevel Methods in Educational Research, University if Edinburgh.

Goldstein, H. (1980) 'Fifteen Thousand Hours: A review of the statistical
procedures', Journal of Child Psychology and Psychiatry, 21, 363-369.

Goldstein, H. (1984) 'The methodology of school comparisons', Oxford
Review of Education, 10, 1, 69-74.

Goldstein, H. (1986) 'Multilevel mixed linear model analysis using
iterative generalized least squares', Biometrika, 73, 1, 43-56.

Gray, J. (1990) 'The quality of schooling: Frameworks for judgement',
British Journal of Educational Studies, 38, 3, 204-223.

Gray, J. & Jesson, D. for the Audit Commission (1991) Two B's or Not...?
Schools' and Colleges' A-level Performance, London: Audit Commission.

Gray, J., Jesson, D. & Tranmer, M. (1993) England and Wales Youth Cohort
Study. Boosting Post-16 Participation in Full-time Education: A Study of
Some Key Factors, Sheffield: Department of Employment, ED Research
Series Youth Cohort Report no. 20.

Inner London Education Authority (1980) School Examination Results
in the ILEA 1978, London: ILEA.

Jencks, C., Smith, M., Ackland, H., Bane, M. J., Cohen, D., Gintis, H.,
Heyns, B. & Michelson, S. (1972) Inequality: A Reassessment of the
Effect of Family and Schooling in America, New York: Basic Books.

Jesson, D. (1992) 'Beyond the league tables', Education, 179, 9, 179-180.

Jesson, D. & Gray, J. (1991) 'Slants on slopes: Using multi-level models
to investigate differential school effectiveness and its impact on
pupils' examination results', School Effectiveness and School
Improvement, 2, 3, 230-247.

Jesson, D., Gray, J. & Sime, N. (1991) Participation, Progress and
Performance in Post-Compulsory Education, Sheffield: Department of
Employment, Rese

Madaus, G. F., Airasian, P. W. & Kellaghan, T. (1980) School Effectiveness:
A Reassessment of the Evidence, New York: McGraw-Hill.

Mortimore, P. (1991) 'The nature and findings of research on school effectiveness in the primary sector', in Riddell S. & Brown, S. (eds.),
op. cit., 9-19.

Nuttall, D. L., Goldstein, H., Prosser, R. & Rasbash, J. (1989) 'Differential
school effectiveness', International Journal of Educational Research,
13, 769-776.

Paterson, L. J. (1989) An Introduction to Multi-level Modelling. Paper presented at the ESRC International Conference on Applications of Multilevel Methods in Educational Research, Edinburgh.

Raudenbush, S. W. & Bryk, A. S. (1986) 'A hierarchical model for studying school effects, Sociology of Education, 59, 1-17.

Rutter, M., Maughan, B., Mortimore, P. & Ouston, J. (1979) Fifteen Thousand Hours: Secondary Schools and their Effects on Children, London: Open Books.

Smith, M. S. (1972) 'Equality of educational opportunity: The basic findings reconsidered', in Mosteller, F. & Moynihan, D. (eds.) On Equality of Educational Opportunity, New York: Vintage Books.

Toffler, A. (1981) The Third Wave, London: Pan Books.

Willms, J. D. (1985) 'The balance thesis: Contextual effects of ability on
pupils' O-grade examination results', Oxford Review of Education, 11, 1, 33-41.