Oxford Review of Education, Vol. 19, No. 4, 1993

A Multilevel Analysis of School Examination Results [1]

HARVEY GOLDSTEIN, JON RASBASH, MIN YANG, GEOFFREY WOODHOUSE, HUIQI PAN, DESMOND NUTTALL & SALLY THOMAS

ABSTRACT *Data on examination results from inner London schools are analysed in relation to intake achievement, pupil gender and school type. The examination achievement, averaged over subjects, is studied as is achievement in the separate subjects of mathematics and English. Multilevel models are fitted, so that the variation between schools can be studied. It is shown that confidence intervals for school 'residuals' or 'effects' are wide, so that few schools can be separated reliably. In particular, no fine rank ordering of schools legitimately can be produced. A bivariate model for mathematics and English examination achievement scores is fitted. The student level variance for both subjects is shown to increase from the lowest to the highest intake achievement group, with moderately high correlation between the subjects. The paper discusses the implications of these findings for the publication of 'league tables' of school examination and test scores.*

INTRODUCTION

There is now a considerable literature on methods for comparing schools and other institutions on the basis of the achievement of their students. The important paper of Aitkin and Longford (1986) established that the minimal requirement for valid institutional comparisons was an analysis based upon individual level data which adjusted for intake differences and used efficient techniques of multilevel modelling. In that paper and the discussion on it, several outstanding problems were raised. The purely technical problems of carrying out the estimation for large data sets have been solved fairly effectively by the development of computer programs, summarised in Kreft *et al.* (1990). The remaining problems are concerned with the existence of suitable measurements which can be used to adjust for intake, and any other relevant differences, the multivariate nature of school outcomes, and the kinds of interpretations which can be made of results. The present paper addresses these latter issues. Specifically it looks at two measures of intake achievement for each school in the study, and examines the interpretational issue by studying the dimensionality of school differences.

DATA

The data are examination results from 5748 students in 66 schools in six Inner London Education Authorities. These students had data on their General Certificate of Secondary Examination (GCSE) grades in mathematics and English, together with a total score for all the subjects taken in that examination. A description of the types of data

TABLE I. *Mean scores by intake categories*

	Mathematics	English	Total	%
VR group 1	0.75	0.80	0.81	24
VR group 2	−0.07	−0.07	−0.07	56
VR group 3	−0.75	−0.80	−0.81	20
Total	0.00	0.00	0.00	100
Corrn with LRT	0.51	0.58	0.58	

and the scoring system used is given in Nuttall *et al.* (1989). For mathematics and English, a scale ranging from 0 (no grade awarded) to 7 (grade A) was used in the analysis and, for the total score, the scale ranged from 0 to 70. These students also had scores on a common reading test taken when they were 11 years old—the London Reading Test (LRT) (Levy & Goldstein, 1984) and were graded also into three categories on the basis of a verbal reasoning (VR) test at 11 years (Nuttall *et al.*, 1989). Table I shows the standardised mean scores on the three outcome measures by verbal reasoning group and the correlations with the standardised reading test score. All three scores are scaled to have mean zero and standard deviation 1. The pattern is similar for all three response variables.

The original number of students on whom some examination data had been obtained was 8857 in 74 schools. Students were omitted from the analysis if they did not have both intake measures. Where students did not take an examination they are given a score of 0, the same as if they obtained an ungraded result. The exclusion of these students resulted in a sample with a higher total examination score, 23.7 as opposed to 20.0. This differential loss of students with lower examination achievements needs to be borne in mind when interpreting the results, and is a persistent problem with data of this kind.

Two separate models have been fitted to the data. The first analyses the total examination score and the second is a bivariate analysis of the English and mathematics scores. All the response variables have been transformed using normal scoring to conform as closely as possible to multivariate normality.

TOTAL EXAMINATION SCORE

The explanatory variables used in this analysis were as follows: standardised London reading test (LRT); verbal reasoning category; gender; school gender (mixed, girls, boys); school religious denomination (State, Church of England, Roman Catholic, other). Formally, the model is written as follows

$$y_{ij} = \sum_{h=0}^{5} \beta_h x_{hij} + \sum_{h=6}^{10} \beta_h x_{hj} + \sum_{h=0}^{2} u_{hj} x_{hij} + \sum_{h=0}^{1} e_{hij} x_{hij} \qquad (1)$$

where i refers to student and j refers to school. Throughout this equation the subscript 0 refers to the constant term (= 1), the subscript 1 to LRT and 2 to the dummy variable for VR group 1. Subscripts 3–5 refer to the square of LRT, the dummy variable for verbal reasoning group 2, and the dummy variable for gender. The subscripts 6–10 refer to the five school level defined variables listed in Table II under the heading 'fixed part'. The first summation refers to the explanatory variables defined at the student level, the second to those defined at the school level, the third to the random part of the model defining variation at the school level, that is level 2, and the fourth summation defines the random variation at the student level, that is level 1. We also have, at level 2,

TABLE II. *Analysis of total examination score*

Fixed Part	Estimate	SE	
Intercept	−0.53		
LRT	0.37	0.02	
LRT2	0.035	0.008	
VR1–VR3	0.70	0.04	
VR2–VR3	0.31	0.03	
Girls–boys	0.13	0.03	
Girls—mixed school	0.07	0.06	
Boys—mixed school	0.09	0.07	
CE—State school	−0.04	0.13	
RC—State school	0.20	0.06	
Other—State school	0.12	0.16	
Random—between schools:	Cov. matrix (corrns) Intercept	LRT	VR1 (VR3,VR2)
Intercept	0.055		
LRT	0.012 (0.75)	0.0046	
VR1 (VR3,VR2)	0.013 (0.40)	0.009 (0.97)	0.019
Random—between students	Intercept	LRT	
Intercept	0.55		
LRT	0.046	0	

The level 1 (between students) variance is thus a linear function of LRT score, given by: variance = 0.55 + 0.092 LRT.See Goldstein (1987) for a discussion of modelling level 1 variation. Likelihood ratio test statistics for: (a) Level 1, LRT (covariance) $\chi^2_1 = 66.0$, $P < 0.001$. (b) Level 2, VR1 (VR2,VR3) variances and covariance $\chi^2_3 = 11.0$, $P = 0.012$. (c) Level 2, LRT variance $\chi^2_3 = 24.7$, $P < 0.001$.

$$var(u_{hj}) = \sigma^2_h,\ var(e_{hij}) = \sigma^2_{eh}.$$

The level 1 contribution to the variance is

$$\sigma^2_{e0} + 2\sigma_{e01}\, x + \sigma^2_{e1}\, x^2$$

That is, a quadratic function of x and the individual level variances and covariance in this expression do not have separate interpretations. On the other hand, since the x_{hij} are defined at level 1, the level 2 variances and covariances are interpreted directly as between-school variances and covariances for the relevant coefficients.

Several exploratory models were fitted and the above table gives estimates for the model found to give the most satisfactory fit.

In the fixed part of the model Table II shows the average effects of the explanatory factors fitted jointly. The effect of school gender is small and the differences are about the same order of magnitude as the estimated standard errors. There seems to be a small advantage for those attending Roman Catholic schools. Girls do better than boys, and as expected there are large differences between those in the different verbal reasoning categories and there is a strong quadratic relationship with LRT. Turning to the between school variation, we see that the relationship between examination score and LRT varies, as does the difference between verbal reasoning categories 1 and 3, with high positive correlations. At the student level the variance increases with increas-

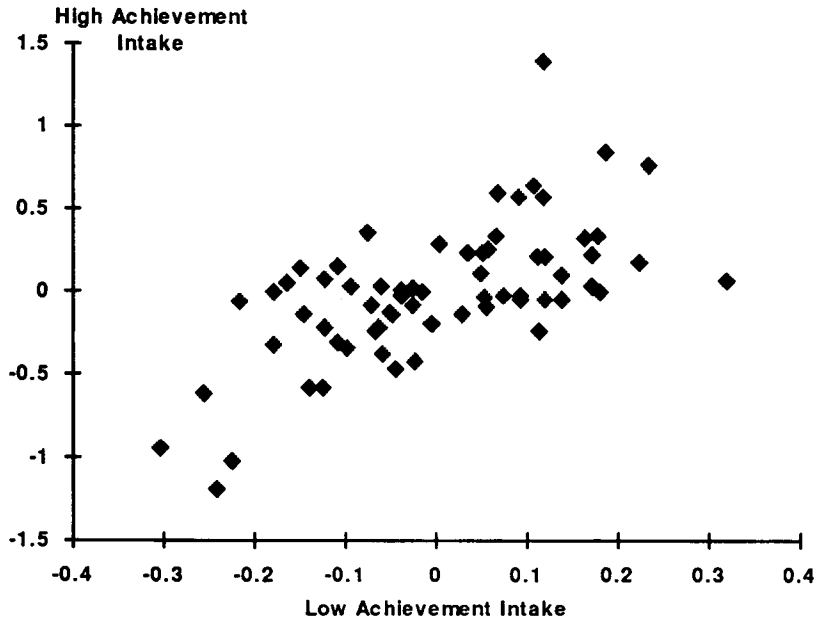

FIG. 1. Total examination score residuals.

ing LRT score, so that the estimated variance for an LRT score of -2 is 0.37 and for one of $+2$ is 0.73. The proper specification of the level one variance is important in order to increase precision and to enable complex variation to be fitted at level 2 and above.

School Residuals

In model (1) the level 2 residuals have a distribution over schools and having fitted the model we can estimate these residuals. These are obtained by estimating the regression model with the (unknown) residuals as responses. The resulting estimates are often known as 'shrunken' estimates since, like all regression predictions they have smaller variances than that of the true values, in this case:

$$\sigma^2_{u0},\ \sigma^2_{u1}.$$

To illustrate the implications of the model, we form particular extreme combinations of the school residuals. For each school we have a value of u_{hj} and we form, using the sample estimates, the two combinations

$$u_{0j} - 2u_{1j},\ u_{0j} + 2u_{1j} + u_{2j}$$

that is, first the estimated school 'effect' for a student with an LRT score of -2, the approximate lower 2.5th percentile, and in verbal reasoning group 2 or 3, and second the estimated 'effect' for a student at the approximate upper 2.5th percentile and in verbal reasoning group 1. These are plotted against each other in Fig. 1.

As pointed out above, there is a positive correlation between the school 'effects' for the low and high achievers on intake. Nevertheless, there are some schools with below average values for the low achievers which have above average values for the high

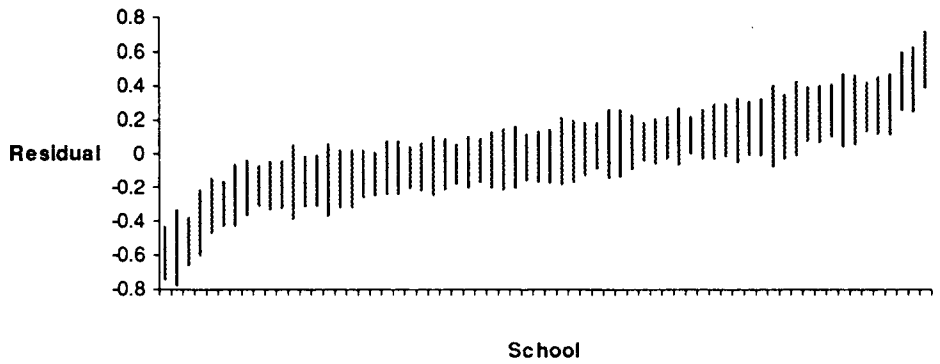

FIG. 2. Total examination score residuals—confidence intervals.

achievers, and vice versa. This emphasises the point that schools appear to be differentially effective for different kinds of students.

Because the residuals are estimated, they have a sampling variation, and this enables us to construct confidence intervals for them. In Fig. 2 are shown approximate 95% confidence intervals for estimate of the intercept residual, that is the school 'effect' estimated at the mean LRT score for those in verbal reasoning groups 2 and 3. It should be noted that these intervals are calculated separately for each residual, and are based upon the estimated standard error which in general will be an underestimate of the true standard error. For comparing any two particular schools, the usual significance test and confidence interval procedures can be used. As can be seen, there is a very considerable overlap of intervals, which suggests that it is not possible statistically to discriminate easily between schools. In particular, there are no natural division points in the sequence of estimates which would allow us to classify schools into homogeneous subgroups. This has important implications for the use of such estimates, as will be discussed later.

JOINT ANALYSIS

We now turn to the analysis of the English and mathematics examination scores. We have chosen these because, in principle, these examinations are taken by all students. It would be possible to carry out a joint analysis of these two scores together with the total score on the other subjects, but for simplicity of interpretation we shall restrict ourselves to just the two. Also for simplicity, we use only the student level variables as explanatory variables, and at the between-school level we use only the intercept and LRT coefficient as random variables.

We specify a multivariate model by treating the multiple variates within each student as the level 1 classification. In this case, therefore, there are two level 1 units within each student (level 2) with schools at level 3. Further details of the model formulation for multivariate data can be found in Goldstein (1987). In the fixed part of this model we see from Table III that the average difference between girls and boys is 0.1 units in favour of the boys for mathematics and 0.23 units in favour of the girls for English. For LRT and verbal reasoning categories, there is little difference between the relationships for maths and English. In the random part of the model the LRT coefficients for English and maths do not vary greatly. The standard deviation for maths is 0.006 units

TABLE III. *Joint analysis of English and mathematics examination scores*

Fixed Part	Estimate	SE
Intercept (Maths)	−0.37	
Intercept (English)	−0.50	
LRT (Maths)	0.021	0.001
LRT (English)	0.027	0.001
LRT2 (Maths)	0.00014	0.000046
LRT2 (English)	0.00018	0.000046
VR1–VR3 (Maths)	0.772	0.041
VR1–VR3 (English)	0.625	0.041
VR2–VR3 (Maths)	0.340	0.027
VR2–VR3 (English)	0.256	0.028
Girls–boys (Maths)	−0.100	0.025
Girls–boys (English)	0.225	0.025

Random—between schools	Int. (Maths)	Int. (English)	LRT (Maths)	LRT (English)
Int. (Maths)	0.037			
Int. (English)	0.012 (0.09)	0.046		
LRT (Maths)	0.0006 (0.49)	0.0004 (0.30)	0.00004	
LRT (Eng)	0.0003 (0.51)	0.00006 (0.09)	0.000006 (0.32)	0.000009

Random—between students	Maths	English
(a) VR1		
Maths	0.63	
English	0.44 (0.73)	0.58
(b) VR2		
Maths	0.52	
English	0.43 (0.83)	0.52
(b) VR3		
Maths	0.28	
English	0.16 (0.53)	0.32

while that for English is only 0.003. While schools differ in terms of overall maths and English performance, at least for English, there is little differential effect according to LRT at intake. The intercepts for maths and English have a small correlation (0.09), and there is only a moderate correlation for the intercepts and LRT coefficients for both maths and English.

At the student level it is clear that the between-students variation decreases from VR1 to VR3 category students. This is similar to the finding in the analysis of total score, where the lower achieving intake students (based on LRT) had smaller variance. We have used verbal reasoning category in this analysis because the results are more clear cut than they are when LRT is used as in the analysis of total score. If a model is fitted with just a variance term for mathematics and English and a covariance term at level 1, the effect is to increase the standard errors for both the fixed part of the model and the level 2 random parameters by up to 20%. The estimates of the coefficients and parameters themselves do not change appreciably, but the decrease in precision emphasises the importance of accurate level 1 modelling.

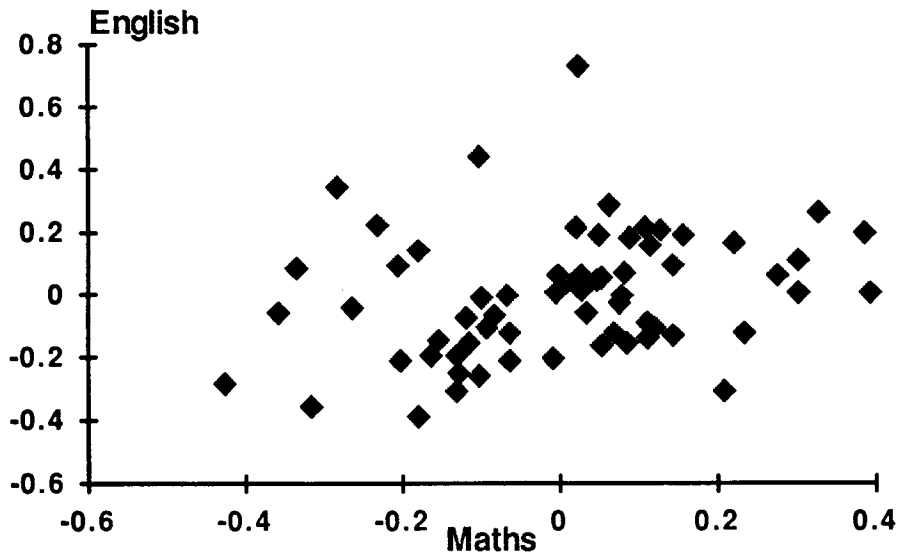

FIG. 3. Intercept residuals.

School Residuals

To illustrate the relationship among the school level residuals or 'effects', Fig. 3 plots the residual estimates for the two intercepts, that is at the mean LRT score. Fig. 3 shows that there is little relationship between English and maths performance. The school with the greatest English residual is only average for maths and one of the schools with high maths residual has a low value for English. This relationship is for those students with average LRT scores. Since the LRT coefficients vary across schools, the relationship between maths and English residuals will also vary with the LRT score.

DISCUSSION

The principal aim of this analysis has been to show how differences between schools in examination results vary by intake achievement and by curriculum subject considered. In addition, the paper explores the extent to which schools can be compared based on residual estimates of 'effectiveness'.

It is clear that there is no single dimension along which schools differ. The ordering of school effects depends on the intake achievements of students as well as the curriculum subject being examined. It is also clear that the uncertainty attached to individual school estimates, at least based upon a single year's data, is such that fine distinctions and detailed rank orderings are statistically invalid. This has important implications for published 'league tables' whether or not these are adjusted for intake achievement and whether or not multilevel modelling has been used. Nevertheless, a study of residuals differentiated by intake achievement and by subject, can suffice as a screening device and as feedback to individual schools about potential problems.

An important feature of the present analysis is the modelling of the between-student, level 1, variation. There is an association between the between-student variance and the

intake achievement score, with increasing variation as the intake achievement increases. This is of substantive interest and it is also important to incorporate it in the model since it helps to ensure that the overall model is correctly specified and will generally improve the precision of the remaining parameters. Furthermore the proper specification of the level 1 variance structure is often necessary to ensure that non-zero estimates of higher level variance structures can be obtained.

In the present analysis no account has been taken of possible unreliability in the LRT score or the VR band allocation. If the reliability is low the estimates may be seriously biased and this issue is explored in a separate paper (Yang *et al.*, 1993). The LRT score has a quoted reliability of 0.95 which is high enough to avoid serious bias (Levy & Goldstein, 1984). Gray *et al.* (1990) found little evidence for random slopes in their own LEA data sets. One explanation may lie in the use of different intake achievement measures, and another explanation may be that Inner London schools are more heterogeneous than those within other LEAs. Further research is needed to clarify this issue.

The use of verbal reasoning and reading achievement measures to adjust for intake is not entirely satisfactory. Ideally, when total examination score is used as the response, initial achievement measures should cover the full range of school subjects taken in the examination. When either mathematics or English is the response, a basic requirement is for the intake achievements to cover these subject domains. This raises the important issue of a common definition of these across ages. It is clear, however, that at least in the case of mathematics, neither the verbal reasoning nor the reading measure satisfy this requirement. Thus the mathematics results may be expected to underestimate the amount of variation explained.

ACKNOWLEDGEMENTS

We are grateful to the London Local Education Authorities for supplying the data and to the Association of Metropolitan Authorities for supporting the project to collect and analyse examination data. The research upon which this paper is based was partly supported by the Economic and Social Research Council (UK) through its funding for the Multilevel Models Project at the Institute of Education.

NOTE

[1] This paper was read to the European Conference on Educational Research, University of Twente, June 1992.

REFERENCES

AITKIN, M. & LONGFORD, N. (1986). Statistical modelling in school effectiveness studies (with discussion), *Journal of the Royal Statistical Society*, A, 149, pp. 1–43.

GOLDSTEIN, H. (1987) *Multilevel Models in Educational and Social Research* (London, Griffin; New York, Oxford University Press).

GRAY, J., JESSON, D. & SIME, N. (1990) Estimating differences in the examination performances of secondary schools in six LEA's: a multilevel approach to school effectiveness, *Oxford Review of Education*, 16, 2, pp. 137–158.

KREFT, J.G.G., DELEEUW, J. & KIM, K.S. (1990) *Comparing Four Different Statistical*

Packages for Hierarchical Linear Regression: GENMOD, HLM, ML3, VARCL (UCLA Centre for research on evaluation, Los Angeles, California, USA).

LEVY, P. & GOLDSTEIN, H. (1984) *Tests in Education* (London, Academic Press).

NUTTALL, D.N., GOLDSTEIN H., PROSSER, R. & RASBASH, J. (1989) Differential school effectiveness, *Journal of Educational Research*, 13, pp. 769–776.

PROSSER, R., RASBASH, J. & GOLDSTEIN, H. (1991) *ML3: Software for Three-Level Analysis*, (London, Institute of Education).

YANG, M., WOODHOUSE, G., PAN H., GOLDSTEIN, H. & RASBASH, J. (1993) Adjusting for measurement unreliability in multilevel modelling (submitted for publication).

Correspondence: All authors: University of London Institute of Education, 20 Bedford Way, London WC1H 0AL, UK.