School Effectiveness and School Improvement
1995, Vol. 6, No. 2, pp. 97-114

0924-3453/95/0602-0097$6.00

A Multi-Level Analysis of School Improvement: Changes in Schools' Performance over Time

John Gray[1], David Jesson[2], Harvey Goldstein[3], Keith Hedger[1], and Jon Rasbash[3]
[1]Homerton College, Cambridge, England
[2]Sheffield University, England
[3]Institute of Education, London University, England

ABSTRACT

The improvement of schools takes place over extended periods of time. Consequently longitudinal studies which track successive cohorts of pupils through their schooling are required if estimates of the extent of improvement are to be established. To date, hardly any studies have collected the necessary data. Those studies which have had appropriate data have tended to emphasise the extent of stability of schools' effectiveness over time rather than the extent of any changes. A shift in conceptual framework is called for if improvements in schools' effectiveness are to be the central focus of concern.

The study is based on three successive cohorts of pupils passing through some 30 English secondary schools. It uses examination results as the outcome measure and includes a prior attainment measure amongst the variables used to control for differences between schools' intakes. A multi-level strategy for conceptualising and modelling data on schools' changes in performance over time is offered.

In common with earlier studies the research shows that there is a good deal of stability in schools' effectiveness from year-to-year; only a small proportion of the schools in the study (between a fifth and a quarter) were improving or deteriorating in terms of their effectiveness. A particularly striking finding of the research was that whilst several schools improved in effectiveness only one initially 'ineffective' school did so consistently. The implications of the study for future research on school improvement are discussed.

INTRODUCTION

There have been many studies of school effectiveness over the past decade and an equally large number of studies of school improvement. In recent years there has been increased interest in bringing the insights of

Correspondence: John Gray, Homerton College, Cambridge CB2 2PH, United Kingdom.

Manuscript submitted: September 28, 1993
Accepted for publication: August 1, 1994

the two approaches together. A number of stumbling blocks have emerged, however, amongst which the divergence of methodological approaches has been one of the most prominent (see Reynolds, Hopkins and Stoll, 1993 for a fuller analysis). In this paper we explore an important aspect of school improvement from the perspective of school effectiveness. How much do schools change in terms of their effectiveness over a number of years? Answers to this sort of question will enable us to secure a better framework for studying the mechanisms of school improvement.

The improvement of schools takes place over extended periods. Estimates by school improvers of the time needed vary, but often spans of three to five years are mentioned. A typical school effectiveness study might follow one cohort through (say) the primary or secondary stage of schooling by collecting evidence on pupils' prior attainment at or around the point of entry as well as on their subsequent performance four or five years later. Researchers have become quite ingenious at building up longitudinal and quasi-longitudinal studies that follow up a cohort of pupils without necessarily waiting the full five years of 'real' time that it takes for them to pass through the system. Some researchers have collected data on a second cohort as well. However, as we shall shortly show, hardly any have been in a position to extend their data-collection to cover a third cohort.

School improvement is essentially about bringing about changes in levels of school performance. One is looking for trends across years. Three years' worth of data is, unfortunately, the minimum requirement for identifying a trend. It is not surprising, therefore, that we have very few studies of school improvement conducted within the school effectiveness paradigm. They simply take too long to conduct.

There are some other reasons as well why we are short of such studies. Probably the most important of these is that school effectiveness researchers have been primarily concerned with the *stability* of school effects rather than their converse. When they have been in a position to collect a second year's data they have mostly sought to replicate their findings. A lack of stability between years has threatened the validity of their research. Instability is, of course, essential for the study of change. An inappropriately limiting conceptual framework may, therefore, have inhibited developments in this area. An orientation towards change is required.

Another reason may be that much of the research evidence reports rather high levels of stability across years. Changes in performance in many of these studies would have appeared to be rather small.

PREVIOUS STUDIES

In all we identified a total of eleven studies which had collected data on more than one cohort of pupils. We judged inter-year correlations of 0.9 to be 'high' and those above to be 'very high'; correlations between 0.5 and 0.9 to be 'middling' ones; and correlations of less than 0.5 to be on the 'low' side.

Four studies looked at British secondary schools. Of these three reported 'high' or, indeed, 'very high' correlations between one year's estimates and the next's (Willms and Raudenbush, 1989; Nuttall et al., 1989; Sime and Gray, 1990) and one reported 'middling' correlations (Rutter et al., 1979). One study of Dutch secondary schools (Bosker and Guldemond, 1991) also estimated 'high' correlations whilst another (Roeleveld, de Jong and Koopman, 1990) reported 'middling' ones.

Correlations in primary school studies appear to have been lower. Mandeville (1988) and Rowan & Denk (1982) in the USA report 'low' to 'middling' correlations whilst Blok and Hoeksma (1993) in the Netherlands also found 'middling' ones.

We were able to identify just one study outside the primary and secondary age-ranges. Tymms and Fitz-Gibbon (1990) report 'low' to 'middling' correlations at the departmental level in the English post-16 sector.

There is one other study which has frequently been mentioned in relation to changes in schools' effectiveness. Teddlie and Stringfield (1993) followed up eight pairs of 'outlier' schools over a number of years. It is not possible, unfortunately, to describe their data in the same way in terms of correlations as for the other studies. They claim, however, that over the period of their study four of their schools were 'stable effective', four were 'stable ineffective', five were 'improving' and three were 'declining'.

We are not completely confident that the eleven studies we have identified are the only ones which have looked at issues of stability and change. We would be surprised, however, if there were more than two or three others.

MINIMUM REQUIREMENTS FOR STUDIES OF SCHOOL IMPROVEMENT

To conduct satisfactory studies of changes in schools' performance over time a number of factors must be brought together. In our view these include:

- measures of outcomes and prior attainment on individual pupils;
- data on a minimum of three cohorts and preferably more;
- a multi-level statistical analysis; and
- an orientation towards examining the data for systematic changes in schools' performance over time.

It is our belief that none of the studies conducted to date have combined all of these four requirements at the same time although two or three were in a position to do so. In this study we provide an initial analysis of some data available to us which lays out the kind of framework we believe is required.

THE DATA

In the present case data were available on three cohorts of pupils aged 16 who took their final compulsory GCSE examinations in the summers of 1990, 1991 and 1992 respectively. The GCSE is a national examination which pupils take at the age of 16-plus in a variety of different subjects; they receive a specific grade in each subject for which they are entered.

The pupils in the study attended over 30 different secondary schools of varying organisational types taking pupils from the ages of 11 to 18 within one LEA. In most key respects the framework for the study was similar to that employed in the Rutter study (1979) and other major British studies of secondary schooling.

The main concern in constructing the data-sets was to match pupils' GCSE results with their prior attainment scores at around the age of 11, shortly prior to entry to secondary school. In 1990 this matching was attempted for a 20% random sample of all pupils in the relevant cohort. In 1991 and 1992 this was increased to all pupils (100%) although, for various reasons to do with pupil movements in and out of the schools, the achieved matched samples were smaller. Complete data were assembled on 7829 pupils. Plans are in hand to repeat the exercise for the 1993 cohort. In addition to data about the school attended, prior attainment and GCSE results, information was collected on each pupil's gender.[1]

Figure 1 shows the distribution of the prior attainment scores in each of the three years. It will be noted that these were approximately normally distributed with a mean of just over 100 and a standard deviation of between 12 and 13 points (see notes to Table 1). Figure 1b provides the cumulative distributions of the prior attainment measures in each of the three years and indicates that they were very similar across the three cohorts.

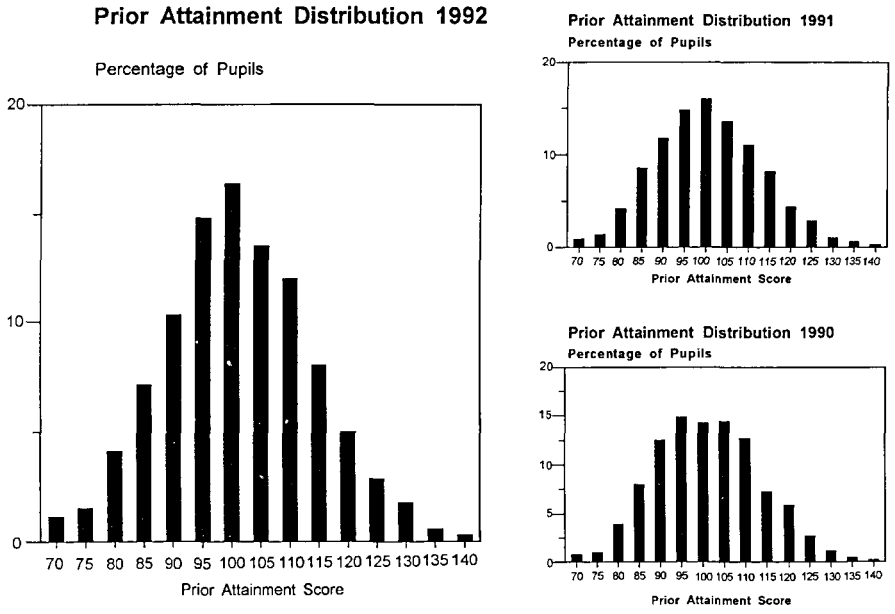

Fig. 1a

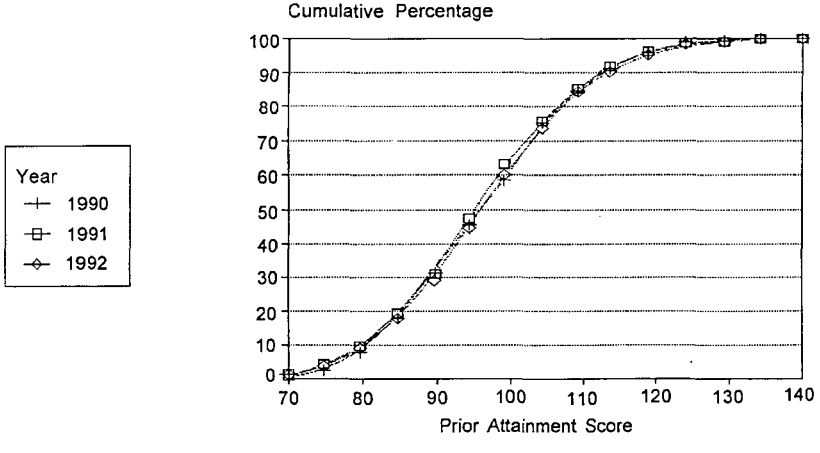

Fig. 1b. Distribution of Prior Attainment Score in each of three years.

Figure 2a shows the distribution of GCSE Examination Scores. For each pupil a total score was calculated based on allocating a score for each grade they obtained in each subject of their Year 11 GCSE examinations. Using a scale which has been frequently employed in recent research on English secondary schools a grade A was given 7 points; a B 6 points; a C 5 points; a D 4 points; an E 3 points; an F 2 points; and a G 1 point. The average GCSE Examination Score per pupil is not, of course, the only measure of outcome that can be constructed from the available data. The average grade across examination subjects and results in, for example, the 'best five' subjects taken could also be considered. However, these alternative measures are not the main focus of the enquiry here.

It will be noted from Figure 2a that these scores diverged somewhat from a normal distribution with rather more zero scores than would have been expected. They were therefore transformed into Nscore distributions across the whole data-set to produce, as far as possible, normally-distributed residuals for the model. This transformation also tends to make the relationship in the data more nearly linear, thus simplifying the model specification. It is these measures which were subsequently employed in the analyses as the dependent variable.

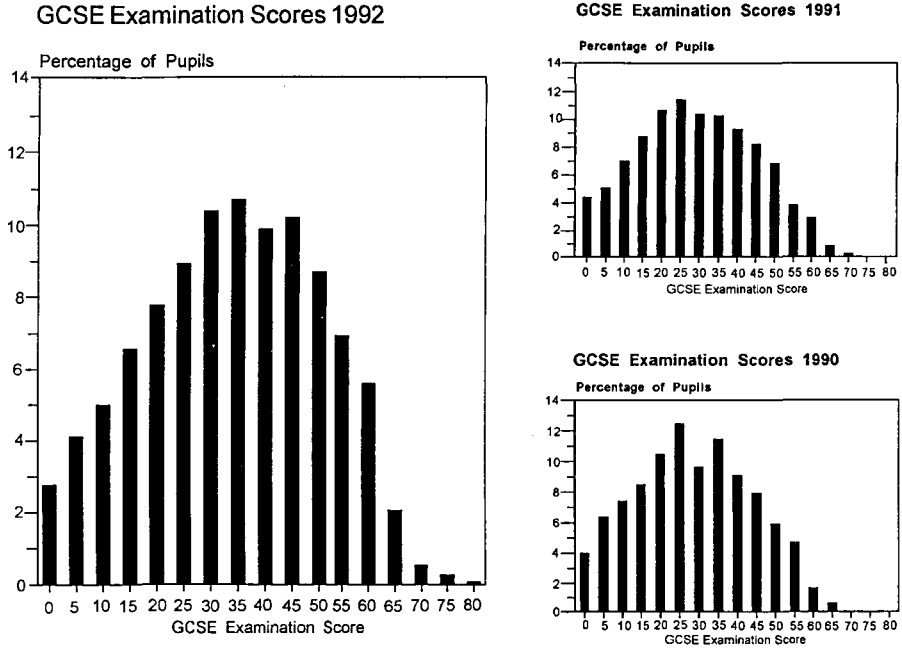

Fig. 2a. Distribution of GCSE examination scores.

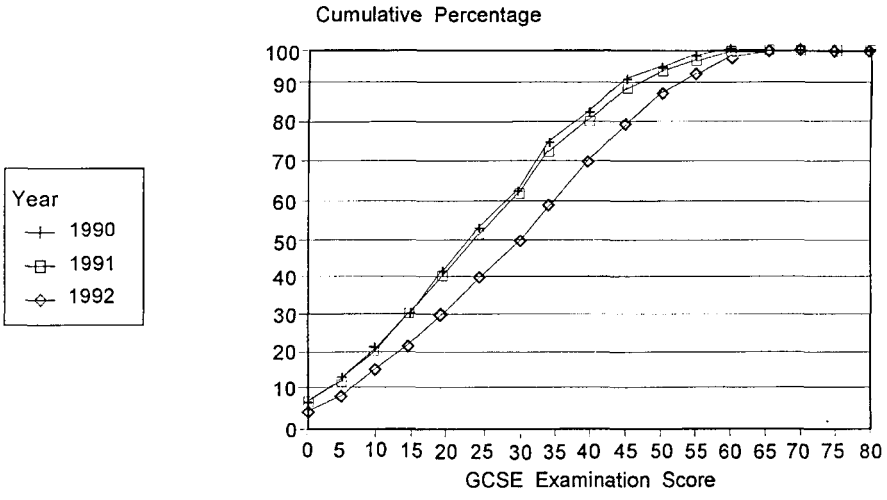

Fig. 2b. Cumulative distribution of GCSE Scores in each of three years.

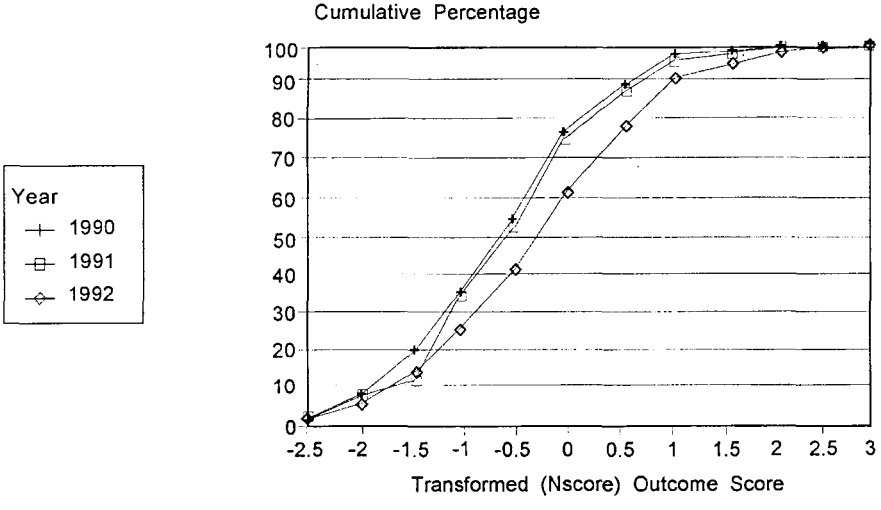

Fig. 2c. Distribution of Outcome Scores in each of three years.

Figure 2b shows the cumulative distributions for the GCSE scores for each of the three cohorts. It is of interest that the distributions for 1990 and 1991 were virtually identical. There was clearly a change, however, between 1991 and 1992. Levels of performance in 1992, in line with national trends, were substantially above those achieved in the two previ-

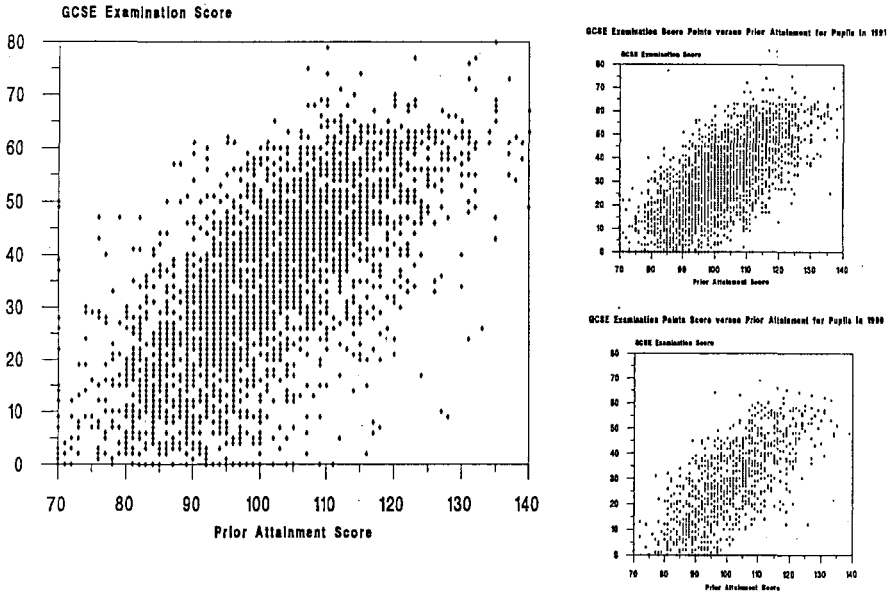

Fig. 3. GCSE Examination Score Points versus Prior Attainment for Pupils in 1992, 1991 and 1990.

ous years. These changes were also largely reflected in the Nscore distributions (see Figure 2c).

The basic relationships between the prior attainments and subsequent GCSE Examination Scores of individual pupils for each of the three cohorts are displayed in Figure 3. These reflected patterns which are already familiar from previous research, prior attainment proving to be a good predictor of subsequent exam performance. The correlation for 1990 was 0.71; for 1991 it was 0.70; and for 1992 0.69.

DESCRIPTION OF THE BASIC MODEL

A series of multi-level analyses were conducted on the data using the ML3E package (see Prosser et al, 1992). In the main the analysis followed the procedures described in an earlier analysis by Goldstein et al. (1993) for the analysis of a single year's data. In the first instance we estimated the model separately for each of the schools. However, this strategy not only failed to take account of the continuity of each school's

existence over time but also failed to incorporate the effect which time as a variable might have on each school's effectiveness. We therefore sought to build a model in which 'time' was explicitly modelled.

The basic model used in the analyses can be written as follows:

$$y_{ij} = \beta_{0j} + \beta_{1j}t_{ij} + \beta_2 v_{ij} + \Sigma\alpha_k x_{kij} + e_{ij}$$

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

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

where 'y' is the Nscore transformation of pupils' GCSE scores; 't' is the year whose coefficient is allowed to vary across schools (the level two units); and 'v' is the prior attainment score. The subscript 'i' refers to the student; 'j' to the school; and 'k' indexes the covariates. The intercept, as is usual, varies across schools. The final summation represents the remaining explanatory variables in the model.

FIXED PART

The Nscore transformation of the GCSE Examination Scores was used as the outcome measure. In the fixed part of the model the main explanatory variables fitted were: prior attainment (measured about the value of 100), gender, year of sitting GCSE examinations and a dummy variable representing the situation in the one selective institution amongst the schools in the data-set. This selective school was retained in the analysis because, in any group of schools, an 'outlier' of this kind is possible. It is, moreover, interesting to ascertain how it differs from the remainder rather than treating it as a 'nuisance' and simply eliminating it. As Table 1 shows, the between-school variation was estimated from the non-selective schools; a similar strategy was employed by Aitkin and Longford (1986) in their pioneering paper.

In addition to using prior attainment as a linear factor this variable was also represented by quadratic, cubic and quartic transformations. All were found to contribute significantly to the resulting model.

Given our interest in modelling changes in schools' performances over time, two 'trend' terms were incorporated to capture the form of any temporal relationships which might be found in the data; both linear and quadratic forms were employed.

Interaction terms representing the joint influence of gender and prior attainment were included in the model. A compositional measure, based on the average prior attainment score in each school, was developed and this was also fitted as an interaction term with pupils' prior attainments.

Separate estimates were made for the selective school by using dummy variables.

RANDOM PART

The basic structure of the data can be represented as a three-level model with pupils grouped into the year of their taking GCSE examinations and then clustered by school. However, we used only two levels for the analyses reported here (pupils and schools) since this provided more detailed information about the variation at each year and changes from year-to-year. In practice, therefore, year was represented by a coded variable which took sequential values in each year.

Given our interest in identifying change over time, a factor representing this was incorporated into the level two random structure. This factor was allowed to vary randomly across schools and sought to capture, as a year-on-year linear slope, the time-related changes occurring in each school's outcome measure (nett of changes in intake or of changes that were common to all schools). In addition, each school's intercept (that is its estimated performance in 1990) was allowed to vary randomly across schools.

The 'fixed' part of the multi-level model expresses the average change over time of the exam score. Thus, the inclusion of a quadratic relationship with year is simply a device for acknowledging that this average may change from year-to-year and thereby allows a trend to be clearly seen. In the 'random' part of the model we have modelled the components of this trend (that is the linear and quadratic terms) to see if these vary across schools. In fact, only the linear term does so in these data. With the availability of data from further years we would expect to be able to fit the quadratic term as random at level 2 as well. In the present case, therefore, the principal estimated (posterior) residuals of interest are those for the intercept and year slope for each school.

Several alternative ways of estimating the model were explored. A summary of the model which was eventually fitted is shown in Table 1. As with all statistical analyses we have attempted to provide an interpretable summary of the relationships. In presenting the results we have left in some strictly non-significant terms (at the 5 per cent level) as we believe these will be of interest to some readers.

Table 1. The Basic Multi-Level Model.

Fixed Coefficients	Estimate (s.e.)
Intercept	–0.326 (0.035)
Prior attainment	0.053 (0.0013)
Gender	0.225 (0.016)
Prior attainment ^ 2	–0.00015 (0.00009)
Prior attainment ^ 3	–0.000012 (0.000002)
Prior attainment ^ 4	0.00000019 (0.0000001)
Gender * Prior attainment	0.004 (0.001)
School X	1.265 (0.289)
School X * Prior attainment	–0.098 (0.025)
School X * Prior attainment ^ 2	0.001 (0.006)
Year	–0.030 (0.042)
Year ^ 2	0.100 (0.017)
Mean Prior attainment for school	0.016 (0.006)
Mean Prior attainment * prior	0.0013 (0.0002)
Random Parameters	
Level 2:	
Intercept	0.0187 (0.007)
Year	0.004 (0.002)
Intercept – Year covariance	–0.002 (0.003)
Level 1:	
Intercept	0.47 (0.01)
School X	–0.165 (0.03)

Notes:
Prior attainment is measured about a score of 100.
Year is coded: 1990 = 0; 1991 = 1; 1992 = 2; and is zero for School X.
Gender is coded: male = 0; female = 1.
Distribution of Prior Attainment: 1990 mean 101.1, s.d. 12.4; 1991 100.6, 12.7; 1992 101.2, 12.8.
GCSE Exam Score: 1990 mean 28.5, s.d. 15.3; 1991 29.2, 15.8; 1992 34.3, 16.8.
Test for Year variation at Level 2; $\chi^2 = 9.3$ (2 d.f.), P = 0.01.
Normal scores are estimated for the response variable using total sample data.
Estimated between school variance at each year:
1990: 0.187; 1991: 0.189; 1992: 0.199.

Not surprisingly, as in many previous studies, prior attainment and gender emerged as important predictors. Other terms in the equation are best seen as being of a more exploratory nature since the patterns to be expected are less firmly established in previous research.

THE EXTENT OF CHANGE

The multi-level model described in Table 1 was used to produce estimates of school-level residuals over the three-year period. It will be noted from the data provided at the foot of the table that between school variance was somewhat greater in the third year than in the first and second years, a pattern which can be anticipated, to some extent, from the 'raw' data presented earlier (see notes to Table 1). For convenience of explication we confine ourselves here to estimates for the median pupil although a number of analyses to explore the extent of differential effectiveness (for which there was some modest evidence) were also conducted.

As previous researchers of secondary school effectiveness in Britain have reported, there was considerable stability from year-to-year. The correlation between the school-level residuals in 1990 and 1991 was 0.94; between 1991 and 1992 it was 0.96; and over the two-year period between 1990 and 1992 it was 0.81. There was, however, also some evidence of changes in performance. Table 2 provides a first fix on the extent of these in relation to changes over the period 1990 to 1992 in a form that can be readily grasped by the non-statistical reader.

Schools' residuals were grouped into the top quarter, the middle half and the bottom quarter for the years 1990 and 1992. Table 2 shows that two out of three schools (68%) were similarly assessed in terms of effectiveness over the period (cells 1, 5 and 9). The table also shows, however,

Table 2. Changes in Schools' Effectiveness Over Time.

	Position in 1992		
Position in 1990	Top Quarter	Middle Half	Bottom Quarter
Top Quarter	(1) 15%	(2) 6%	(3) 0%
Middle Half	(4) 9%	(5) 35%	(6) 9%
Bottom Quarter	(7) 0%	(8) 9%	(9) 18%

Notes: 15% of the schools in the study had school-level residuals in 1990 and 1992 which placed them in the top quarter in both years (cell 1).
The percentages in the table sum to 101%.

that one in three schools (33%) experienced some changes (cells 2, 4, 6 and 8); of these 18% 'improved' (cells 4 and 8) whilst 15% 'deteriorated' (cells 2 and 6). Interestingly, no schools in the study moved dramatically from one position to another over the period of the study (cells 3 and 7).

The structure of the variance between schools was very similar between 1990 and 1991 (see notes to Table 1 for the estimates). However, there was some evidence of greater variation in 1992 which indicates some widening of the differences between schools.

Whilst Table 2 provides some tentative evidence of changes it is based on a rule-of-thumb framework. Figure 4 shows the position in greater detail for all the schools in the study and highlights those schools where the change slopes over time were significantly different from zero (1.4 times their standard error). The four heavy lines which slope up from left to right each represent a particular school whose results have been 'improving' over the three years of the study; the three heavy lines which slope downwards from left to right each represent a school which was 'deteriorating'. Only one school which was deemed to be 'ineffective' at the beginning of the study had consistently improved its performance.

Since we only had three years' data available the analysis has *assumed that these were linear trends*. To talk of non-linearity with just three points in the 'trend' would be to run the risk of over-interpreting the data, but equally we cannot rule the possibility of non-linearity out completely. In practice, there may have been somewhat bigger changes in one year than in the other. In other words, Figure 4 *could* be drawn to emphasise, for example, that there had been a big change in one year and a much smaller change in the next and so on. It is, we believe, convenient to treat them at the moment as if they were linear for heuristic purposes. However, with the addition of data from further years in the analysis the probable complexity of the change process in individual schools can be revealed. Few schools are likely, in practice, to make equal-sized changes in effectiveness year-on-year. The inclusion of non-linear terms in the model allows various possibilities regarding trends in the data to be explored over the longer-term.

Figure 4 also shows the school-level residuals for all the other schools in the sample. It should be noted that these estimates are drawn from the random part of the model, each school having an intercept and a slope, both with a mean of zero. It will be seen that some other schools remained consistently more 'effective' or 'ineffective' than those we have chosen to highlight but were less subject to change.

In substantive terms pupils in the 'improving' schools were gaining an average of around one additional GCSE Examination Score point per year

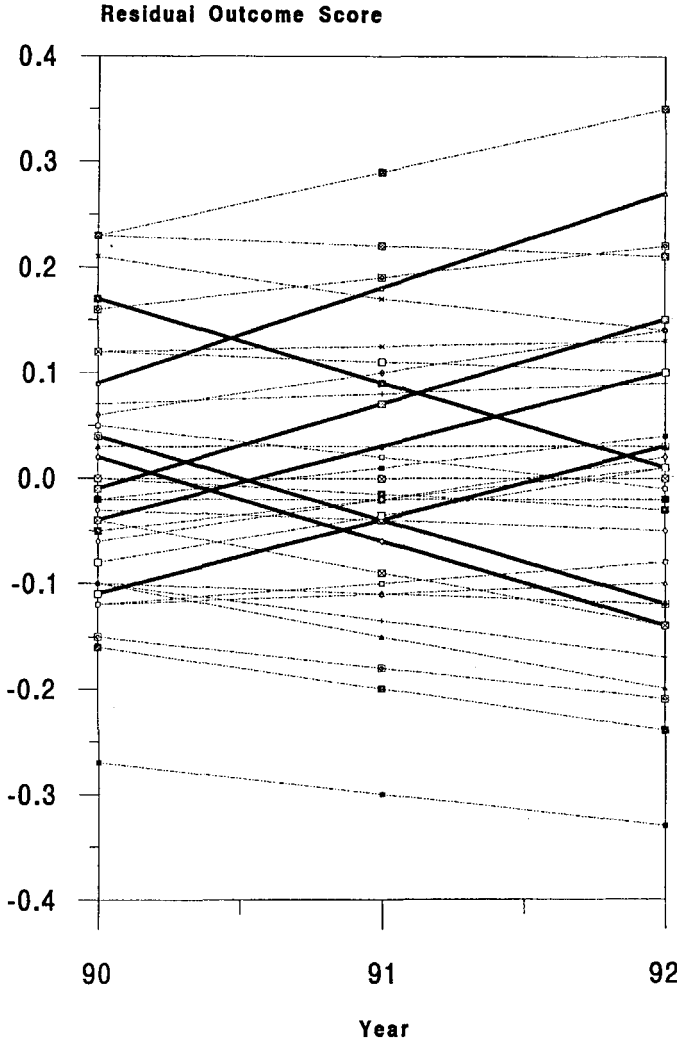

Fig. 4. Significant changes in modelled estimates of school residuals over a three year period.

(equivalent to, for example, an improvement for each pupil from a grade D to a grade C in one subject in each consecutive year and representing about one-sixth of an individual standard deviation on the GCSE Examination Score measure).

DIFFERENTIAL SLOPES WITH RESPECT TO PRIOR ATTAINMENT

In an initial exploratory analysis we utilised a three-level model reflecting the structure of the data (as outlined above). There was some evidence of a differential relationship between outcome score and prior attainment amongst schools (level 3) but the year component (level 2) of schools' performance was poorly estimated (details not reported here).

In subsequent analyses, whilst not discounting the possibility of differential slopes on prior attainment, we pursued the issue of differential rates of change in performance across time using a two-level model. It is clearly possible, theoretically, to incorporate both factors into our models. However, we have preferred to defer that issue until we have at least one further year's data.

CONCLUDING COMMENTS

We have argued that to research school improvement from the perspective of school effectiveness some reorientation is required. To date there have been very few studies indeed which have attempted to do this so the possibilities of comparing our findings with those of previous research are distinctly limited.

First, the evidence available to us suggests that, at any one time, only a small proportion of the schools in any particular locality will be improving or deteriorating in terms of their effectiveness in ways which are substantively significant. Our estimates suggest that this will be between a fifth and a quarter of the schools. Obviously some of these will be improving and others deteriorating. The number of schools which are actually changing at any one time is, however, likely to be small in comparison with the number that may appear to be changing for reasons that are not necessarily connected with changes in their effectiveness such as higher-attaining pupil intakes and boosts to pupil outcomes that are common to large numbers of schools. Only one 'ineffective' school in the entire sample was found to be improving consistently over the period of the study.

Second, changes in effectiveness at any one time (either in the form of improvement or deterioration) are likely to appear modest, at least in comparison with the differences in overall effectiveness prevailing at any one time. Researchers need to be sensitive to the sizes of the changes that appear to be occurring and to their potential significance in educational

terms. Consistent improvements in a school's performance over a five-year period could, for example, turn a relatively ineffective school into a relatively effective one.

Third, with only three years' data it is premature to talk of longer term trends; every additional year's data represents a potentially significant enhancement to our understanding. In some further analyses, not reported here, we explored the patterns of change which were taking place. Some schools appeared to start off slowly and then accelerate; others did the reverse. Although we have presented the change process as if it were a simple linear process we think it unlikely that this is the only form that change can take. Reports from the field lead us to anticipate rough and uneven developments. From the point of view of understanding them initially, however, it is probably easier to smooth them out but this is just a heuristic device. Each school will have its own 'natural history' of change to which particular changes in effectiveness need to be related.

Fourth, the extent to which changes in effectiveness are dependent on the outcome measures employed needs to be considered. There are different ways of constructing measures from the examination data available which require further exploration; improvements may, for example, be concentrated in some subjects and not others, or result from pupils being entered for more examinations in total (see Gray, Jesson and Jones, 1986). Whilst most of these exam measures are highly correlated a total score may give different results from one based on the average grade in each subject entered or the results obtained in the 'best five' or 'best seven' subjects and so on. At the same time it needs to be recognised that improvements in examination performance may not go hand-in-hand with other kinds of change.

Fifth, and in the meantime, we obviously need to start asking what is causing particular schools to improve or deteriorate relative to others. We suspect that the current school effectiveness literature, when revisited, may prove to be dominated by concepts relating to the correlates of effectiveness rather than the correlates of improvement (i.e. changes in effectiveness). Back in the early eighties Purkey and Smith (1983) asked some important questions about the state of existing knowledge. 'Were different strategies required', they asked, 'for low-achieving schools to raise their scores than for high-achieving schools that (were) beginning to decline?' What was needed to maintain a school's success 'once it was deemed academically effective?' And how did 'different improvement strategies affect sub-populations of pupils in a school?' In this research framework 'ineffective' schools which are improving are at least as interesting as research sites as 'effective' schools which maintain their positions.

A decade later the answers that can be provided to these questions, on the basis of research on school effectiveness, are still somewhat tentative. Merely exhorting less effective schools to copy the current practices of more effective ones may not work. Clearly, with the development of research designs covering longer time-periods, a new stage of research enquiry is in prospect.

NOTE

1. A socio-economic background measure (free school meals) was also available but was not employed in the analyses reported here as it was only collected for the second and third cohorts. It will, however, be used more systematically in future analyses and we anticipate that it will make a small additional contribution to the explanatory model.

ACKNOWLEDGEMENTS

A number of people have given generously of their time to help us identify previous studies of relevance to our current concerns. We should particularly like to thank: Sally Tomlinson, Bert Creemers, Peter Mortimore, Carol Fitz-Gibbon, Charles Teddlie, David Reynolds, Roel Bosker and Doug Willms. We should also like to note the help offered by two anonymous referees in critiquing the basic model.

REFERENCES

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

Blok, H., & Hoeksma, J.B. (1993). The stability of school effects over time: an analysis based on the final test of primary education. *Tijdschrift voor Onderwijsresearch, 18* (6),331-342.

Bosker, R.J., & Guldemond, H. (1991). Interdependency of performance indicators: an empirical study in a categorical school system. In S. Raudenbush, and J.D. Willms (Eds.), *Schools, Classrooms and Pupils.* New York: Academic Press.

Goldstein, H., Rasbash, J., Yang, M., Woodhouse, G., Pan, H., Nuttall, D., & Thomas, S. (1993). A multi-level analysis of school examination results. *Oxford Review of Education, 19* (4), 425-433.

Gray, J., Jesson, D., & Jones, B. (1986). The search for a fairer way of comparing schools' examination results. *Research Papers in Education, 1* (2), 91-122.

Mandeville, G.K. (1988). School effectiveness indicators revisited: cross-year stability. *Journal of Educational Measurement, 25,* 349-366.

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

Prosser, R. Rasbash, J., & Goldstein, H. (1992). *MLE3: Software for Three-Level Analysis.* London: Institute of Education.

Purkey, S.C., & Smith, M.S. (1983). Effective schools: a review. *Elementary School Journal, 83* (4), 427-452.

Reynolds, D., Hopkins, D., & Stoll, L. (1993). Linking school effectiveness and school improvement. *School Effectiveness and School Improvement, 4* (1), 37-58.

Roeleveld, J., de Jong, U., & Koopman, P. (1990). Stabiliteit van schooleffecten. *Tijdschrift voor Onderwijsresearch, 15* (5), 301-316.

Rowan, B., & Denk, C.E. (1982). Modelling the academic performance of schools using longitudinal data: an analysis of school effectiveness measures and school and principal effects on school-level achievement. *Instructional Management Program Publication.* San Francisco: Far West Laboratories.

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

Sime, N., & Gray, J. (1991). *Struggling for improvement: some estimates of the contribution of school effects over time.* Paper presented to the Annual Conference of the British Educational Research Association.

Teddlie, C., & Stringfield, S. (1993). *Schools Make a Difference: Lessons Learned From a Ten-Year Study of School Effects.* New York: Teachers College Press.

Tymms, P., & Fitz-Gibbon, C.T. (1990). *The stability of school effectiveness indicators.* Paper presented to the Annual Conference of British Educational Research Association.

Willms, J.D., & Raudenbush, S. (1989). A longitudinal hierarchical linear model for estimating school effects and their stability. *Journal of Educational Measurement, 26,* 290-32.