CHAPTER 6

DIFFERENTIAL SCHOOL EFFECTIVENESS

DESMOND L. NUTTALL,* HARVEY GOLDSTEIN,† ROBERT PROSSER† and JON RASBASH†

*Inner London Education Authority, London, U.K.
†Institute of Education, University of London, London, U.K.

Abstract

Studies of school effectiveness are briefly reviewed, pointing to the need to study effectiveness for sub-groups within each school as well as overall. The results of a multilevel analysis of a large dataset covering the years 1985, 1986 and 1987 and using examination performance as the outcome measure are presented, revealing substantial differences between ethnic groups. The findings also show that the effectiveness of a school varies along several dimensions, and that there is also variation over time. The implications of these findings are discussed.

Introduction

This chapter describes preliminary analyses of a large dataset held about secondary schools in inner London. It explores an issue of great concern to policy-makers, teachers and parents, namely whether some schools are more effective than others not only in a general sense, which is now well established in the literature, but also in the sense of being equally effective for all groups (e.g., boys and girls, ethnic minority groups and so on). In this chapter, the effectiveness of schools is measured in terms of their students' success in public examinations at age 16. The Inner London Education Authority, currently the largest education authority (school district) in the U.K., with 140 secondary schools, serves an inner city, multi-racial community as well as more affluent suburbs; the differences between the schools it controls are very marked in terms of their social and ethnic composition. Its twin aims are to improve the quality of education, and to ensure equality of opportunity.

The issue of measuring and describing the effectiveness of schools has become even more significant in England and Wales after the passing of the 1988 Education Reform Act. Among other things, this Act requires each school to publish a wide range of performance measures based on a series of tests and assessments across the curriculum (TGAT, 1988a, b), as well as a diverse set of performance indicators as a result of the devolution of financial control to the schools. The proposals require that these indicators should be published without statistical adjustment to reflect the different characteristics of

the intakes of the schools, though the publication of adjusted results in addition is not prohibited. Fair and comprehensible ways of presenting performance indicators in context, and to reveal differences between sub-groups of students, are therefore urgently needed.

Background

School effectiveness has been extensively studied, and the two most influential studies in England are Rutter, Maughan, Mortimore, and Ouston (1979) and Mortimore, Sammons, Stoll, Lewis, and Ecob (1988), for secondary schools and primary schools respectively. Both used schools in the area of ILEA. The ILEA has also analysed school effectiveness routinely, and published ratings of the effectiveness of all its secondary schools on two occasions (ILEA Research and Statistics, 1986, 1987a). 'Effectiveness' is taken to be the difference between the actual 'output' of the school and the 'output' expected (in the statistical sense) of a school with identical student characteristics. The measure of output used was the aggregated public examination results of students aged 16. A variety of measures of the characteristics of pupils attending each school were investigated using aggregate level multiple regression analyses; three factors consistently emerged as significant in analyses over the years. First was the proportion of the age group in each school eligible for free school meals (a measure of economic deprivation); the second was the proportion of pupils in VR Band 1, which is a London-wide measure of performance at age 11 generating norms of 25% of the population in VR Band 1, 50% in VR Band 2 and 25% in VR Band 3 over the ILEA area as a whole (allocation of individuals to each band is carried out by the head and teachers at each primary school on the basis of their judgement of the pupils' overall attainment; the *number* in each band in each school is assigned by the ILEA on the basis of the *number* of pupils scoring on each quartile on a test of verbal reasoning administered throughout the Authority). The third factor was the proportion of girls in the age group in each school, reflecting the fact that nearly half the secondary schools in London are for girls only or boys only, leading to sex imbalance in many of the co-educational schools, as more parents prefer single-sex education for their daughters than for their sons.

The more recent of the two routine reports (ILEA Research and Statistics, 1987a) noted a number of significant limitations to the analyses and the ILEA agreed to suspend the publication of ratings of effectiveness for each secondary school while the methodology was improved. The most significant limitation was the use of aggregated data for each school, rather than data on each individual student, e.g., his or her sex. Woodhouse and Goldstein (1988) have shown the dangers of relying on aggregated data in the context of the analysis of examination results for local education authorities in the U.K., and have advocated the use of multilevel models (Goldstein, 1987). The methodological literature now shows that these models are universally preferred in the study of school effectiveness (see, for example, Aitkin & Longford, 1986; Bryk & Raudenbush, 1987; Mortimore *et al.*, 1988).

Two major issues concern methodological investigators in the field of school effectiveness at the present time. The first issue is the stability of effects across time. The second issue is of greater social and educational significance. Recent studies (e.g., Hallinger & Murphy, 1986; Cuttance, 1988a; Teddlie, Stringfield, Wimpleberg, & Kirby,

1989) show that the characteristics of effective schools are not necessarily the same for schools in areas of very different socio-economic status, while the work of Gray, Jesson and Jones (1986), and Cuttance (1988b) shows that some schools are more effective with some sub-groups (e.g., those of high attainment on entry) than with others (e.g., those of low attainment on entry).

The ILEA Junior School Project (Mortimore *et al.*, 1988) found that some schools were more effective with particular sub-groups than other schools. For example:

> These results suggest that, in general, schools which had a positive effect in promoting reading progress for one sex also tended to have a positive effect for the other, whilst those which were ineffective for one sex were likely to be ineffective for the other. Nonetheless, there was some variability. Although for the majority of schools (29) effects for girls and for boys were in the same direction (positive for both or negative for both) the results were discrepant in 12 schools. In eight of these schools, effects on reading progress were positive for boys, but negative for girls. (Mortimore *et al.*, 1988, p.210.)

This chapter presents a more detailed investigation of these issues, for ILEA secondary schools rather than junior schools; it might be expected that secondary schools, as much larger organisations, could show greater diversity in their differential effectiveness. Moreover, because of the larger numbers of students in secondary schools and because many more schools are included in the present sample, better estimates of sub-group differences can be made.

Method

The dataset comprises the results of public examinations taken at, or about the age of 16 for three cohorts of ILEA students, those attaining the age of 16 in the school years ending in 1985, 1986 and 1987 respectively. These examination results are readily available and have to be published in aggregated form by each school as a requirement of the 1980 Education Act.

The examination results were made available for each student within each secondary school by the examining bodies. Schools were asked to provide information on the sex, VR band on entry to secondary school (as described above), and the ethnic background on each student in the cohort. Two basic ethnic categories were used: 'Black' and 'White'. Within 'Black', the sub-categories were: African, African-Asian (in 1985 only), Arab, Bangladeshi, Caribbean, Indian, Pakistani, South-East Asian, and other Black; the 'White' sub-categories were: Irish, English/Scottish/Welsh (although in 1985 those two categories were merged), Greek, Turkish, Other European White and Other Non-European White. Some groups were small, and the results for such groups are not always shown in the analyses discussed below. Full information about the samples in 1985 and 1986 are given in ILEA Research and Statistics (1987b), and the 1987 sample was very similar. Not all 140 secondary schools provided student-based data on ethnicity in each year; the number of schools where complete data are held for all three years is 64, but all 140 provided data in at least one year. Certain other information about the schools has also been incorporated into the analyses; for example whether each school is mixed or single sex, and whether it is fully maintained by ILEA (i.e., a county school) or voluntary (i.e., supported by the Church of England or the Roman Catholic Church). Additional information about the schools, including aggregated data about the student body, is

available from other databases and will be included in more extensive analyses still being carried out.

The data were analysed with multilevel modelling software (Rasbash, Prosser, & Goldstein, 1989), using three levels: between students, within schools/between years, and between schools.

Results

The multilevel model provides estimates both of the fixed or average effects, such as the difference in the performance of boys and girls overall, and of the random effects, such as the variation in the boy–girl difference across co-educational schools. The model can also provide estimates of such differences for each school (Goldstein, 1987, Chapters 2 and 3). Table 6.1 provides estimates of the fixed effects in the three-level analysis (using data from three consecutive years involving 140 schools and 31,623 students).

Apart from the constant term (or 'intercept' estimate), the estimates refer to differences between groups: for example, the difference between the performance of girls and boys is 2.5 score points in favour of girls, with a standard error of 0.2. The largest differences are

Table 6.1
Fixed Effect Estimates

Coefficient	Estimate	Standard error
Constant (intercept)	17.8	—
Girls minus boys	2.5	0.2
VR Band 1 minus Band 3	19.0	0.3
VR Band 2 minus Band 3	8.2	0.2
Ethnic group[a]		
African	4.0	0.5
Arab	4.4	1.1
Bangladeshi	4.7	0.7
Caribbean	−0.4	0.2
Greek	4.6	0.7
Indian	7.3	0.5
Pakistani	6.0	0.6
SE Asian	8.3	0.6
Turkish	3.7	0.4
Other	3.8	0.4
Year	1.4	0.2
Boys schools minus mixed schools	0.8	0.3
Girls schools minus mixed schools	1.4	0.3
Church of England schools minus county	1.2	0.4
Roman Catholic schools minus county	2.4	0.3
Free school meals (FSM) proportion	−0.41	0.04
FSM percentage squared	0.003	0.0004

[a]Each ethnic group is contrasted with the English, Scottish and Welsh group.

between students in the three VR bands. It should be noted that a score of 7 points is awarded to a Grade A in the GCE O-level examination, so that the average difference of 19 points between students in VR Bands 1 and 3 amounts to nearly 3 Grade As. The performance of the different ethnic groups is in each case compared with that of the students of English, Scottish, Welsh and (in 1985 only) Irish backgrounds (ESWI), who form the largest single group. All ethnic groups perform significantly better than the ESWIs, except those of Caribbean background who perform slightly but not significantly worse.

The relationship between examination score and time is 1.4 points increase per year which is statistically significant. The contrasts between single sex and mixed schools are also statistically significant, in favour of single sex schools. Voluntary denominational schools' examination performance is significantly better than that of county schools, especially for the Roman Catholic schools.

The final two coefficients describe the relationship between the examination score and the proportion of the 16-year-olds in each school who were eligible for free school meals. Thus the average score difference between students in schools where 10% of 16-year-olds are eligible for free school meals and those with 30% eligible is about 6 points.

In the case of all these coefficients it should be remembered that the estimates are not estimates of the actual differences (say, between the performance of girls and boys) but the difference *after* taking the other factors into account, in particular after adjusting for VR band at intake. The differences therefore reflect progress made during secondary schooling.

Table 6.2 shows the random effects, that is, the variation between students and within schools and, at level 3, the extent to which the differences between the sub-groups (as shown in Table 6.1) vary between schools, and relate to each other: more technically, the variances and covariances of the differences. There are potentially very many of these difference parameters and not all of them can be fitted in a single model. Moreover, there are only small correlations between the 'year trend' coefficient and the other coefficients that vary across schools, and it was not therefore necessary to fit these covariances; hence the appearance of 'parameter not fitted' in the bottom row.

In Table 6.2 it can be seen that differences in the performance of VR Band 1 and Band 3 students vary substantially between schools: these differences have a variance of 17.4 (and therefore a standard deviation of 4.2), around an average of 19.0 (from Table 6.1). So in some schools the difference is as small as 11 points and others as large as 28. The difference in the performance of VR Band 2 and Band 3 students however has a variance of only 2.8 (i.e., a standard deviation of 1.7) around a mean of 8.2. The sex difference (in mixed schools) has a standard deviation of 1.4 around a mean of 2.5, implying that there are a few schools where boys actually do better than girls. The difference in the performance of Caribbean and ESW students has a standard deviation of about 1 score point, around a mean of -0.4.

Discussion

The differences shown in Table 6.1, for example between the sexes and between different kinds of schools, are well established and come as no surprise. The differences between ethnic groups have also been reported earlier (ILEA Research and Statistics,

Table 6.2
Random Parameter Estimates

Between students (level 1) variance	96.7
Within schools, between years (level 2) variance	1.2
Between schools (level 3)	

	Intercept	VR1/VR3	VR2/VR3	Sex	Caribbean/ESW	Year
Intercept	2.9					
VR1/VR3	−1.9 (0.0)	17.4				
VR2/VR3	−0.1 (0.0)	6.1 (0.9)	2.8			
Sex[a]	−1.5 (−0.6)	2.4 (0.4)	0.6 (0.2)	2.1		
Caribbean/ESW	−0.4 (−0.2)	−1.8 (−0.4)	−0.5 (−0.3)	−0.3 (−0.2)	1.1	
Year	0.1 (0.0)	NF[b]	NF	NF	NF	0.5

Note: Variances and covariances, with correlations shown below covariances and in brackets.
[a]The sex coefficient varies only across mixed schools.
[b]NF = parameter not fitted.

1987b) both before adjustment and after adjustment for other factors. They are of great concern to the ILEA and are being investigated by schools, teachers and inspectors. An important limitation of the present analyses is that data about the socio-economic level of the students' families are unavailable. It should be recognised that the ESWI population of inner London is not representative, socially or economically, of the total ESWI population of the United Kingdom and Eire, and there could be confounding of the ethnic differences with socio-economic differences.

One kind of effect shown in Table 6.1 is of particular interest, exemplified by the proportion of the 16-year-old age group in each school eligible for free school meals. This kind of effect is termed a *compositional* effect and it has often been hypothesised that, over and above the expected differences in performance attributable to differences between individuals attending each school, greater *concentrations* of underperforming groups will further depress performance (or vice versa). The finding that the performance declines as the proportion of students eligible for free school meals increases is a potential example of such a compositional effect (but it is not a true example because data were only available at the aggregate level for this variable and not at the individual level as well). (A few preliminary analyses within a single year (Nuttall, 1989), using proportions in VR Bands 1 and 3, and the proportion of students of Caribbean background, have shown no significant compositional effects, but this topic is being studied further.)

The results in Table 6.2 are of great interest and bear out the hypothesis that schools' performance varies along several dimensions associated with sub-groups, some schools narrowing the gap between boys and girls or between students of high and low attainment on entry, and some widening the gap, relatively speaking. Furthermore, other analyses

(not reported here in detail) indicate that other ethnic group differences vary across schools, even more than the Caribbean–ESW difference. For example, the Pakistani–ESW difference has a standard deviation of some 3 score points across schools. It is, of course, those schools that narrow the gap by raising the performance of the lower achieving group (rather than by lowering the performance of the higher achieving group) that are of special interest. It would be valuable to study such schools in depth in cooperation with expert observers, such as ILEA inspectors, to explore possible reasons for their differential performance.

As with compositional effects, the number of possible differences that could be explored (e.g., all the ethnic group differences) is too large to include sensibly within one model. Further study of the statistical and practical significance of particular differences and compositional effects is being carried out.

The stability of the efforts over time is an important consideration. The standard deviation of the year trend coefficient is 0.7, indicating that the performance of some schools increases by about 5 points, and that of others not at all over the period 1985 to 1987. In addition there is an unexplained between-year standard deviation of about 1 score point. The reasons for this may be partly to do with an inadequate statistical adjustment. In particular, attainment on entry is available only in the three broad VR bands; it is being replaced by a continuous variable in 1988 (derived from a reading comprehension test). The lack of stability may also be partly to do with the unreliability and lack of comparability of the examination scores. This analysis nevertheless gives rise to a note of caution about any study of school effectiveness that relies on measures of outcome in just a single year, or of just a single cohort of students. Long time series are essential for a proper study of stability over time.

Finally, the relatively small correlations in Table 6.2 should be noted. For example, the school difference between the performance of VR Band 1 and 3 students is not very strongly correlated with the sex or Caribbean–ESW difference. Furthermore, the Band 1–3 differences are virtually uncorrelated with the intercept, i.e., the effect for the students in VR Band 3. This implies that knowing which schools widen or decrease the Band 1–3 *difference* tells one nothing about whether those schools do well or badly for those in VR Band 3. There is, however, a relatively small variation for the intercept which means that the variability in performance of VR Band 3 students between schools is small; hence the variability of VR Band 1 students' performance between schools must be substantial.

A number of the further analyses and investigations that are required have been referred to above. Other further analyses will investigate sensitivity to the specification of the model employed and to the scaling of the outcome variables, as previous work (Goldstein, 1987) has drawn attention to the variation in results as a consequence of variation in model specification. It is also hoped to use the results of examinations in specific subjects (e.g., in English or in mathematics), rather than a composite outcome measure, to explore differential effectiveness in different parts of the curriculum, as Mortimore *et al.* (1988) using an even wider range of outcome measures (both cognitive and non-cognitive) found evidence of differential effectiveness.

Concluding Comments

In summary, this research has found that school effectiveness varies in terms of the

relative performance of different sub-groups. To attempt to summarize school differences, even after adjusting for intake, sex and ethnic background of the students and fixed characteristics of the schools, in a single quantity is misleading.

The findings are not consistent with those of Smith and Tomlinson (1989) who argue that, because they found that the overall variation between schools in examination performance was much greater than the variations in the differences between ethnic groups, it is appropriate to conceive of a single dimension of school effectiveness. Our research indicates, with three years' data in 140 schools as opposed to one year's data in 20 schools, that it is more meaningful to describe differences between schools for different sub-groups: the concept of overall effectiveness is not useful.

Finally, we wish to stress that the implicit definition of school effectiveness in terms of examination performance used here is limited, since examinations represent only a partial, albeit important, measure of 'output'. The results cannot necessarily be generalized to other measures.

Acknowledgements — We would like to thank colleagues in ILEA Research and Statistics Branch, particularly Steve Greenhill and Sona Chumun, for their help in preparing the data for analysis.

References

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

Bryk, A. S., & Raudenbush, S. (1987). Application of the hierarchical linear model to assessing change. *Psychological Bulletin*, **101**, 147–158.

Cuttance, P. (1988a). Intra-system variation in the effectiveness of schools. *Research Papers in Education*, **3** (3), 180–216.

Cuttance, P. (1988b). *Modelling variation in the effectiveness of schooling.* Edinburgh: Centre for Educational Sociology.

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.

Goldstein, H. (1987). *Multilevel models in educational and social research.* London: Charles Griffin.

Hallinger, P., & Murphy, J. F. (1986). The social context of effective schools. *American Journal of Education*, May, 328–355.

ILEA Research and Statistics (1986). *Looking at school performance*, (RS 1058/86). London: ILEA Research and Statistics.

ILEA Research and Statistics (1987a). *Actual and predicted examination scores in schools*, (RS 1129/87). London: ILEA Research and Statistics.

ILEA Research and Statistics (1987b). *Ethnic background and examination results — 1985 and 1986*, (RS 1120/87). London: ILEA Research and Statistics.

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

Nuttall, D. L. (1989). Differential school effectiveness. Paper given at the American Educational Research Association Annual Meeting, April 1989, San Francisco, USA.

Rasbash, J., Prosser, R., & Goldstein, H. (1989). *ML2: Software for two-level analysis (three-level upgrade).* London: Institute of Education.

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

Smith, D. J., & Tomlinson, S. (1989). *The school effect: A study of multi-racial comprehensives.* London: Policy Studies Institute.

Task Group on Assessment and Testing (1988a). *A report.* London and Cardiff: Department of Education and Science and the Welsh Office.

Task Group on Assessment and Testing (1988b). *Three supplementary reports.* London and Cardiff: Department of Education and Science and the Welsh Office.

Teddlie, C., Stringfield, S., Wimpleberg, R., & Kirby, P. (1989). Contextual differences in models for effective schooling in the USA. Paper given at the Second International Congress for School Effectiveness, January 1989, Rotterdam, The Netherlands.

Woodhouse, G., & Goldstein, H. (1988). Educational performance indicators and LEA league tables. *Oxford Review of Education*, **14** (3), 301–320.