Award no: H53627504395

Award Holders: Dr. Sally Thomas

Title: Optimal multilevel models of school effectiveness : comparative analysis across regions

Full Report of Research Activities and Results

ESRC Full Report of Research Activities and Results

Optimal Multilevel Models of School Effectiveness: Comparative Analyses Across Regions

Project H53627504395

Dr Sally Thomas

Background

School effectiveness and the related areas of school improvement and evaluation have been topics for an increasing body of academic study since the 1960's (Coleman *et al* 1966, Jenks *et al* 1972, Rutter *et al* 1979, Mortimore *et al* 1988, Reynolds *et al*, 1996a, Hopkins, Ainscow & West 1994, Sammons, Thomas & Mortimore 1997, Gray *et al* 1999). In contrast, only in the last decade or so have policy makers focused their attention on the possibilities for improving educational practice and pupil performance via more systematic approaches to evaluation and accountability (Reynolds *et al* 1996b). Teachers and Local Education Authorities (LEAs) are now required to use performance data to inform their own evaluations of the education they provide (DfEE, 1996, 1998a; OFSTED, 1998). At the national level, the Office for Standards in Education (OFSTED) inspection reports and school league tables continue to be published as a mechanism for educational accountability. In 1998, the Department for Education and Employment (DfEE) completed a pilot 'value added' study as a supplement to league tables (SCAA 1997; DfEE 1998b).

However, in part, these policy developments have been informed by a relatively small number of quantitative research studies, often employing limited or incomplete datasets. For example, there is little research on value added secondary school effects across different pupil outcomes and regional contexts. Previous work has focused mainly on academic outcomes for a specific curriculum stage (Thomas *et at* 1997a) and few studies have looked at comparisons between regions which vary in terms of both educational policy, socioeconomic and other regional factors (Scheerens & Bosker, 1997 Creemers *et al*, 1994). Also very few studies have investigated methodological developments of the value added approach such as the influence of primary schools on secondary school performance (Goldstein & Sammons 1997) or the extent of differences in effectiveness between classes (Rowe & Hill 1994, Hill & Goldstein 1998). There is a need to develop this area of research in the UK and to clarify the findings of school effectiveness studies in the wider regional and national context using appropriate multilevel techniques. Moreover, recent DfEE and OFSTED reviews of educational research have emphasised that few quantitative studies are reported and recommend that further research is required that replicates and builds on the findings of previous work (Tooley & Darby 1998, Hillage *et al* 1998).

Further background to the study is provided in the nominated paper *Dimensions of Secondary School Effectiveness: corn paring the findings from four academic studies* (Thomas & Smees, 1998).

Objectives

The study aims to replicate, clarify and extend previous research concerning the definition and measurement of secondary school effectiveness in the UK (using multilevel techniques) across a range of outcomes and regional contexts. The objectives were:

(1) To establish the optimal multilevel model(s) for measuring school effectiveness over time using a value added approach in a range of different pupil outcomes (academic and attitudinal[1]). Addressed by research questions 1-4.

(2) To compare the optimal model(s) across different regional contexts (inner city, county LEAs) and education systems within the UK (England, Scotland) and also abroad (Netherlands[2]). Addressed by research question 5.

(3) To identify and define the dimension(s) of school effectiveness that encompass a range of different outcomes and take into account different regional, socio-economic and educational policy contexts. Addressed by research question 6.

The objectives have not changed since the original proposal was submitted and they were met by addressing six research questions described under Methods and Results. The implications of the findings will be discussed in relation to secondary school evaluation in the UK.

Methods

Research Design

The study involved a comparison of secondary school effects drawn from a variety of geographical regions The overall focus of the analyses is on pupil outcomes at the end of statutory schooling (eg GCSE and attitudes at key stage 4), however, additional analyses also examine pupil outcomes at Key stage 3 and post 16.

The methodology adopted a ‘value added’ approach which adjusts for ‘intake’ and aims to separate and measure the school effect and that of other external factors (such pupil prior attainment and socio-economic status) on pupil performance. This approach is well established and further details of the rationale can be found elsewhere (see Nuttall *et at* 1989, McPherson 1992, Fitz-Gibbon 1995, Mortimore, Sammons & Thomas 1994, Thomas, Pan & Goldstein 1994, Sammons, Mortimore & Thomas 1996, Thomas & Mortimore 1996, Goldstein *et at* 1993, Goldstein 1997, Gray 1993, Saunders, 1998). However, it is important to acknowledge at this point that it is impossible to provide statistical adjustment for *all* factors outside the control of the school which have a significant impact on pupil

performance. One limitation of the study is that the outcome and explanatory data employed were pre-selected by the data providers, although these measures have been shown to be important in previous research (Sammons *et al* 1994). Therefore, the value added effects represent the school effect *and other effects not accounted for in the analysis.* The aim is to identify the optimal multilevel model given the data available.

Samples and Data

Six datasets relating to a variety of regions in the UK and abroad (Lancashire, London, Jersey, Scotland, Netherlands, England) were employed for the study (see Appendix 1). Where available the datasets include individual pupil outcomes in different areas (academic and attitudinal), and results for different cohorts and curriculum stages. The following categories of explanatory variables are also included: prior attainment (or attitude) data, background factors (eg entitlement to free school meals (FSM - a measure of low family income), ethnicity, gender and age) and school context (percentage of low attaining pupils - drawn from approximately the bottom 25% ability band).

Statistical Analysis

To establish the optimal model(s) of secondary school effectiveness (objective 1) the statistical technique of multilevel modelling has been employed (Paterson & Goldstein 1991, Rasbash & Woodhouse 1995, 1998). The optimal model is identified in terms of purpose (ie intended use of results), statistical criteria (ie goodness of fit, statistical significance) and appropriateness (ie employing valid outcome and explanatory variables).

The analyses have been carried out in two stages (see Appendix 2). The first stage involved identifying which explanatory variables should be included in the optimal multilevel model(s) over time. The second stage involved extending the optimal model(s) by employing a fixed set of explanatory variables and different model specifications to examine the school residuals for different outcomes, groups of pupils, cohorts, and curriculum stages. The impact on school effectiveness of controlling for effects at other levels of the education system (classroom, region or previous school) was also examined.

Where data were available the analyses were repeated for each regional dataset and the results compared (objective 2). By drawing together the findings from different datasets the aim was to identify and define the *underlying* dimension(s) of school effectiveness (objective 3).

Results

1 Which explanatory variables should be controlled for in the optimal multilevel model(s)?

Academic Outcomes

Overall the findings suggest that in terms of statistical criteria and appropriateness the explanatory variables included in the optimal model may vary slightly for different outcomes as well as different regions (eg school context may or may not be a significant factor). However, for the purpose of employing a consistent set of explanatory variables

across different regional datasets and subject outcomes - as in this study - the optimal model takes into account all prior attainment, background and school context factors. In fact, on average, this model explains the highest percentage of total (52.7%) and school *(77.3%)* level variance in pupil outcomes across datasets in comparison to all other models tested (see Appendix 2).

After controlling for all explanatory factors there are still clear differences in the extent of variation across schools for different outcomes. For example, for multiple cohort datasets, the average percentage of variance in pupil's total score outcomes attributable to schools is 7.6% (2.1% is attributable to differences between cohorts). Using considerable larger datasets, these findings replicate and confirm the results of previous studies. However, the extent of remaining differences between schools varies across regions and this issue is addressed by research question 5. In terms of educational policy these results provide strong evidence of the impact of schools and teachers on pupils' academic outcomes and illustrates the need to provide schools with feedback data on their value added performance.

Attitude Outcomes

The analysis of pupil attitudes is the focus of the nominated paper: *Valuing Pupil Views in Scotland* (Thomas *et al* 1999). The key results are summarised below.

In contrast to the multilevel results for academic outcomes, the data from one pupil cohort (1997) in Scotland shows that secondary pupils' previous attitudes, background characteristics and school context were not particularly good in explaining their later attitudes (on average, the total and school percentages of variance explained were 19% and 55%). Nevertheless, in terms of statistical criteria, the optimal model for all attitude outcomes included pupils' previous attitudes and background characteristics.

Overall the findings indicate that differences between schools in pupils' attitudes are small in comparison to the results for academic outcomes (less than 5% of the total variance in both raw and value added attitude scores is attributable to school). Nevertheless, for policy-makers and practitioners, the findings indicate the kind of affective outcomes that may be most useful to secondary schools for the purpose of providing self evaluation feedback. For example, measures that reflect pupil liking for school (ie engagement scale) and the positive interaction between teachers and pupils in the classroom (ie teacher support scale). However, further research is needed to confirm the results using data from more than one cohort and this work is continuing in collaboration with Lancashire LEA.

2. What **outcomes should be employed in the optimal multilevel model(s)?**

Both separate and multivariate approaches were employed to examine school effects in different academic outcomes (eg total score, language, mathematics, science). Replicating and extending previous research (eg Thomas *et al* 1997a) the results show that schools can have quite different effects in different departments and point to the existence of an

effectiveness dimension for each academic subject. Moreover, the consistency of schools departmental effectiveness across datasets can vary suggesting that whole school policies may have a greater impact in some regions (eg Lancashire) than in others (eg London).

The correlations between schools' adjusted residuals for academic and attitudinal outcomes indicate that the relationship between schools performance in these two areas is relatively weak (range in r: -0.38-0.19, Scotland; -0.31-0.11, Lancashire). However, as may be expected, the 'raw' unadjusted pupil level correlations are somewhat stronger (range in r: -0.27-0.61, Scotland; -0.06-0.38, Lancashire). Educationally important, these new findings support earlier work at the primary level (Mortimore *et al,* 1988) and tentatively suggest that separate dimensions of effectiveness can be identified reflecting different aspects of how schools and teachers can influence pupils' attitudes and achievements.

3. What evidence is there to suggest that the optimal multilevel model(s) should be extended to reflect school effectiveness for different pupil groups, cohorts or curriculum stages?

The optimal models were extended to identify, if possible, differential effectiveness for different pupil groups. However, this approach is only employed for academic outcomes as insufficient data were available to extend the attitude outcome models. The rationale was to examine the consistency of schools overall and departmental residuals for different pupil groups (categorised by individual background factors, cohorts and curriculum stages) in order to identify whether different effectiveness dimensions exist.

(I) Individual background factors

Replicating and extending the findings of previous work (Thomas *et ali* 1999h) the correlations between school residuals for different pupil groups (categorised by prior attainment, gender and FSM) indicate that some differences exist within schools in terms of school and departmental effects for particular groups (see Appendix 3). The results show that non perfect correlations have been found for pupil groups categorised by prior attainment and FSM across four regional datasets with average correlations (r) of 0.59 (prior attainment) and 0.91 (FSM). In contrast, the evidence for differential effects according to gender is weak with average correlations (r) of 0.97. Interestingly, the consistency of school and departmental effects for different pupil groups categorised by prior attainment appears to be stronger in some regions (eg the Netherlands) than in others (eg London).

To examine the FSM differential results in more detail a further model has been employed which includes only prior attainment explanatory variables and therefore does not make the assumption that average attainment differences exist between particular groups. Using this approach, for example in Lancashire, only 22 percent of schools obtain positive value added scores, on average, for pupils entitled to free school meals (FSM), whereas 72 percent of schools obtain positive scores for non FSM pupils. Moreover, in nearly all schools (93%) FSM pupils make less progress on average than other pupils. In spite of the crudeness of the FSM indicator, these findings could be usefully interpreted as most schools having

different levels of effectiveness for pupils who are more (or less) advantaged economically (and possibly also more/less effective learners). Nevertheless, some schools do appear to be able to decrease the attainment gap between more and less advantaged pupils. For the purpose of monitoring equal opportunities and pupil entitlement this evidence points to schools need for feedback that makes explicit the absolute levels of *progress* made by different pupil groups.

(ii) Cohorts

Four multilevel methods of identifying trends over time were examined using the London and Lancashire datasets (see Appendix 4). For each method the correlations between school residuals for consecutive pupil cohorts were calculated to identify differences in school and departmental effects over time. It was found that the stability or instability of school residuals for individual cohorts varied depending on which analysis method was employed. Therefore, different methods are appropriate according to the intended purpose and use of the results. For example, using a separate analysis for each cohort (method 1) or separate intercepts for each cohort in a joint analysis (method 2) emphasises instability over time and these methods are appropriate for the purpose of examining in detail the improvement (or decline) in value added scores over time. In contrast value added results that reflect linear trends (method 3) or the average results of two or more consecutive cohorts (method 4) are more stable over time (ie do not fluctuate randomly from year to year) and are appropriate to examine long term patterns of school performance. Indeed, method (4) has been usefully employed by schools in Lancashire LEA since 1996 as a kind of 'rolling average' of school performance (Thomas, 1998).

Interestingly, comparing the results in Lancashire (over 5 cohorts) and London (over 3 cohorts) to equivalent work by Gray, Goldstein & Jesson (1996) shows that differences between schools in value added time trends are either not statistically significant or much smaller than previously reported (having already accounted for average time trends across schools). That is, irrespective of schools' apparent improvement in raw league table performance, few schools are able to improve substantially in their effectiveness *relative to that of other schools.*

Additional models were also employed to examine whether average trends over time in school effectiveness results varied for different groups of pupils. The findings indicate that the average time trends for particular groups (according to prior attainment and gender) appear to be slightly different. Thus further new evidence is provided that schools need to monitor equal opportunities over time.

(iii) Curriculum Stages

New and previously unreported correlations between schools' effectiveness estimates for key stage 3 and key stages 3 and 4 combined were calculated using 1997 Lancashire data to identify differences in school and departmental effects for different curriculum stages. The non perfect correlations ($r = 0.50$, total score; 0.27, English; 0.65, mathematics; 0.63, science) show that some schools can obtain quite different value added scores according to whether the whole or only part of the secondary curriculum is examined. The results suggest the existence of separate effectiveness dimensions for different National

Curriculum stages, particularly for English outcomes. For government policy makers these findings are important given the publication of a sample of schools value added results for key stage 4 only (DfEE, 1998b). In the light of the current findings a school could appear to be doing well at key stage 4, but not so well at key stage 3 or across both key stages 3 and 4. Overall the results indicate that separate value added measures of effectiveness should be feed back to schools for each key stage as well as for the whole period of secondary schooling.

4. What evidence is there to suggest that the optimal multilevel model(s) should be extended to incorporate additional hierarchical groupings (or levels) within the education system (ie classrooms, regions or previous school)?

Following on from the analyses addressing research question 3 the optimal models for each academic outcome measure were again extended to establish the need to take account of effects attributable to levels other than the school level within the education system. The significance of other levels in accounting for pupil attainment and relative progress is crucial evidence in identifying more precisely the dimensions and appropriate methods of calculating school effects.

(i) Previous Schools

Using 1997 Lancashire data the value added results from cross-classified multilevel models (Goldstein 1995) show that the total and school level variance as well as the variation in pupil outcomes attributable to secondary schools is only slightly reduced (or similar) when allowance is made for primary school attended. Primary school attended does have a statistically significant (at 0.05 level) impact on pupils' GCSE outcomes, but the total variance in value added scores attributed to the primary level is small (ranging from 1%-3%) compared to the secondary level (ranging from 6-12%). These results are in contrast to previous similar work in the UK by Goldstein & Sammons (1997) which reports larger primary school effects than secondary school effects. Interestingly, the findings are in line with similar results in the Netherlands (Snijders & Bosker, 1999) and this suggests that the Goldstein findings may not be representative of the general pattern in UK schools due possibly to the small and incomplete sample employed.

The correlations between school residuals with and without controls for primary school attended are very high ($r > = 0.99$). This indicates that, for the practical purpose of secondary school feedback, little difference in interpretation may be drawn from value added measures which do, or do not, make allowance for primary school attended. Nevertheless, for the purpose identifying overall patterns in and underlying influences on secondary pupils' attainment further work using cross-classified models is required to examine the impact of primary school effectiveness on later achievements by linking *progress* made by pupils at the primary level to subsequent *progress* at secondary school. This approach is also required to examine the related issue of pupils moving school *within* the primary or secondary phase of schooling.

(ii) Classrooms

Using English and mathematics outcome data from the 1997 Scottish dataset the value added results show that total and school level variance as well as the variation in pupil outcomes attributable to secondary school is only slightly reduced when allowance is

made for variation between classes as well as between schools. Nevertheless, some statistically significant (at 0.05 level) variation in pupil performance can be attributed to differences between classes (or teachers) and the impact is considerably greater for mathematics than English outcomes. For English the impact of classroom differences on the estimates of school effects is small for both raw outcomes and in terms of value added (percentage of variance attributable to the class level is 2.2% and 1.8% respectively). In contrast, for mathematics, the variation in raw outcomes attributed to class differences is larger (23.7%) and comparable to previous research where classroom effects on mathematics have been estimated for Scotland (Scheerens, Vermeulen & Pelgrum 1989, Scheerens & Bosker 1997). However, the variation in mathematics outcomes attributed to classes is reduced considerably for value added measures (5.3%). These findings point to the influence of setting in classroom organisation which is more prevalent in mathematics than English teaching. Nevertheless, for mathematics in particular, these results suggest that within school differences in terms of classroom (or teacher) effectiveness may need to be addressed.

Similar to the results adjusting for primary school level, the correlations between school residuals with and without controls for classroom level are very high ($r >= 0.99$). Interestingly, these results suggest that for the practical purpose of providing departmental (or subject) feedback little difference in interpretation may be drawn from value added estimates which do, or do not, make allowance for classroom grouping at least in Scotland. However, further research is required with larger datasets and regions to tease apart the impact on pupil attainment of classroom (or teacher) effectiveness and other school or contextual factors (such as setting, classroom organisation, or teacher mobility).

(iii) Regions

Employing small area definitions of region (involving approximately 10 or fewer secondary schools) no statistically significant regional effects were identified in the Lancashire or London datasets. However, findings from an additional dataset comprising the 1994-95 A/As level results for the whole of England was used to examine the impact of regional effects categorised by LEA. The results show that at the post 16 level small differences in effectiveness between LEAs do appear to exist (percentage of variance attributable to the LEA level was less than 3% for all outcomes tested). It is unclear whether this conclusion would change for secondary results at key stages 3 and 4, although a replication of this analysis will be possible in the future once national curriculum assessments become available nationally.

Importantly, the above results use a single model for England as a whole which controls for the average differences in pupil attainment (according to prior attainment and other factors) across all regions. In research question 5 a meta analysis approach is employed by examining the regional differences via the separate model results for each regional dataset *(cf* Bosker & Witziers, 1995).

5. Are different optimal multilevel model(s) required to reflect regional differences?

The evidence reported for research questions 1-4 show there is some variability in results across regional datasets. However, some differences observed may be due to differences in the availability or definition of explanatory variables included in the optimal models. Therefore, in order to provide clearer evidence on regional differences the results of models which employ only one explanatory variable - pupils' prior language attainment - have also been examined.

The findings show that regions vary in terms of the goodness of fit of models using only language prior attainment as an explanatory factor. However, some regions are more alike than others with Lancashire, Jersey and Scotland obtaining fairly similar results and a much better fit of the data (average percentage total variance explained ranges from 41% to 45%) in contrast to the equivalent results for London and the Netherlands (24%-32%). Moreover, the average percentage of school level variance explained varies across regions (from 51% to 68%). These differences are larger than would be expected on the basis of differences in the method of language assessment between regions and suggest that certain aspects of regional policy or context may be linked to a weaker relationship between pupils' prior and later attainment. For example, London as an inner city area has additional factors of pupil and community poverty as well as long established equal opportunity policies that may result in pupils' outcomes not being predicted particularly well on the basis of their previous reading ability.

The size of school effects also varies considerably between regions. However, as noted above, some regions obtain more similar results than others. Lancashire, London and Scotland - all within the UK - obtain similar variations between schools in terms of effectiveness with differences of approximately 0.8 standard score units between the most and least effective schools[7]. In contrast, Jersey and the Netherlands appear to have much larger variations in schools' effectiveness with pupil outcome differences of more than 1.3 standard score units between the most and least effective schools. The percentage of total variance in value added scores attributable to schools mirrors the regional differences in the size of school effects. Regions again appear to fall into two groups, the first group (Lancashire, London and Scotland) obtain results showing 6-9% of the variance in pupil outcomes is attributable to schools. In comparison, the second group (Jersey, the Netherlands) show a much greater impact of schools (22-24% of variance attributable to schools).

In order to illuminate these findings contextual information about each region has also been examined in terms of (I) the extent of selection amongst schools (ie standard deviation across schools of the percentage of low attaining pupils in each school) and (ii) school differences in raw unadjusted outcomes (ie standard deviation across schools of raw residuals). Interestingly, the regional differences in terms of context reflect the differences observed in terms of the size and impact of school effects. It appears that regions with

larger differences between schools in raw scores and the extent of selection (Jersey, the Netherlands) also show larger differences in the size and impact of school effects.

These findings provide new evidence of regional differences in school effects. Therefore, it seems likely that regional, as well as national, context and policies do have an influence on schools' effectiveness, although a causal link has not been tested. In other words regional context (such as socio-economic or geographical factors) or education policy (such as the extent of selection or private schooling) may limit the possibilities of a school being more or less effective as well as enhance (or inhibit) the overall differences between schools. Moreover, regional differences in context or policy may also be reflected in how well pupil's previous attainment can predict their later attainment. The results highlight the need for further analyses of the size and impact of schools effects in different UK regions, particularly those that vary greatly in terms of selection policy (such as LEAs with a system of selective grammar schools).

6. Does the evidence suggests different dimension(s) of secondary school effectiveness exist and if so how should these be defined?

The findings from research questions 1-3 suggest that at least four dimensions of secondary school effectiveness can be defined specifically in terms of different outcomes, pupil groups, pupil cohorts and curriculum stages. To clarify these definitions the impact or constraints on schools' effectiveness associated with other levels within the education system have also been examined. The findings from research question 4 tentatively suggest that pupils' attendance at a particular primary school, or the classroom grouping of pupils (or at the post 16 level, the location of the institution in a particular LEA), has a relatively small impact in practical terms on the measurement of how effective a school is.

Nevertheless, effectiveness at different levels of the education system (eg individual pupils; classrooms; departments; whole school; regionally and nationally) as well as the interaction between levels need to be continually monitored in order to inform policy development and map out the *boundaries* of school effectiveness. The evidence from research question 5 of regional differences in the range and extent of school effects points to the *interpretation* of schools' effectiveness being regionally dependent and highlights the important role of regional or policy context when defining or evaluating school performance.

Conclusions

Overall, the study findings are of theoretical as well as practical importance pointing to the possible existence of regional sub-systems in the overall functioning of schooling *(cf.* Scheerens & Bosker, 1997). However, further work is needed to investigate and explain the relationship between specific aspects of regional context or policy and the average levels as well as the range of schools' effectiveness in different regions. This approach is in line with other researchers who have emphasised the importance of examining in detail the socio-economic context of pupils and schools in relation to their effectiveness (Mortimore & Whitty 1997, Thrupp 1997, Lauder, Jamieson & Wikeley 1998).

Replicating and extending previous work, the results emphasise the need for school staff and external evaluators to analyse value added data in a sensitive and detailed way. To measure schools' 'value added' pupil attainment measures are needed at the beginning and end of each curriculum stage as well as other background data. A valid framework for secondary school evaluation in the UK needs to incorporate at least four *underlying* dimensions of effectiveness (in terms of different outcomes, pupil groups, cohorts and curriculum stages) and also needs to contextualise the results with regional information. Indeed, separate LEA analyses of school performance may be needed in addition to analyses at the national level. To operationalise this kind of evaluation the data, methods and their limitations need to be transparent and well understood by the evaluators (Goldstein & Myers, 1996).

Hopefully, evidence of this kind will stimulate and inform teachers' evaluation of their own educational practices and capacity for improvement as well as make transparent the constraints, boundaries and context of schools' effectiveness. However, the findings also point to the need for further research on the existence of additional dimensions of school effectiveness in terms of vocational and other pupil outcomes valued by society. Indeed, a comprehensive value added framework for school evaluation might also encompass measures related to numerous other aspects of a school's goals, processes and outcomes (MacBeath & Mortimore 1994, Thomas *et al* 1998).