EDUCATIONAL VALUE ADDED

AND

PROGRAMME EVALUATION

by

DAVID MAYSTON*

University of York

*** The author is Professor of Public Sector Economics, Finance and Accountancy, and Director of the Centre for Performance Evaluation and Resource Management, at the University of York.**

CONTENTS

EXECUTIVE SUMMARY

The evaluation of educational initiatives, such as the Academies programme, raises interesting questions as to the appropriate analytical tools and methodologies to be adopted in the evaluation. In this Research Report, we examine two main literatures which are relevant to this area. The first is the large and growing literature on the assessment of value added in the education sector. The second is the developing literature on programme evaluation in contexts where the conditions for carrying out **randomised control trials** (RCTs) are not fulfilled. We will examine both the opportunities and remaining problems that these approaches present for the evaluation of educational initiatives. In addition, we will seek to bring these two approaches productively together in providing appropriate analytical tools for the evaluation of educational initiatives, whether in a local, national or international context.

The concept of educational value added, and its potential roles in promoting and assessing school effectiveness, are examined in Section 2. Section 3 examines a number of formulations of the value added model, including the use of **multilevel modelling** to take into account the **hierarchical structure** of pupils being educated within classes within schools. While the efficiency of the parameter estimates can be improved by adjusting for the **heteroscedasticity** which such a hierarchical structure implies, there remain potential problems of **instability in the parameter estimates** which the associated Iterative Generalised Least Squares (IGLS) estimation procedure generates. The magnitude of the estimates of the **school effect** which multilevel models generate is examined in Section 4. While estimates of the difference that the school makes range from 1.5 per cent to 25 per cent of the variance explained, most estimates are concentrated around 10 per cent. This itself suggests that differences between schools in general do not make dramatic differences to educational outcomes, once other variables are taken into account. By far the largest part of the variance of individual pupil achievement is that explained by individual pupil prior attainment, underlining the need to make use of value added measures of pupil progress which adjust for this factor.

Section 5 examines a number of multivariate approaches to analysing value added. While the use of Ordinary Least Squares (OLS) multivariate regression analysis on pupil-level data does not produce minimum variance estimates of the relevant parameters, it does nevertheless produce estimates of the fixed effect coefficients of the model that will be **unbiased**. The under-estimate of the standard error of the coefficient estimates which OLS produces,

compared to the use of multilevel error component models, may, however, result in some fixed effect variables being accepted as statistically significant under OLS which would not be found to be so under multilevel modelling. The extent of this under-estimation will tend to be greater, the greater is the **intra-school correlation** between the value-added residuals, and the greater is the number of pupils per school.

However, while the estimates of the fixed effect parameters are unbiased under both multilevel modelling and OLS, the estimates of the **school random effects** that indicate individual school contributions to pupil value added are not unbiased under multilevel modelling. Instead the use of a '**shrinkage factor**' in the multilevel estimation procedure to reflect the assumed reliability of parameter estimates reduces the estimates of school effectiveness, particularly for small schools. Simulation studies show that the use of multilevel models are then not always strongly preferable to the use of OLS on pupil-level data, particularly once possible instabilities in the Iterative Generalised Least Squares (IGLS) procedure of multilevel modelling are taken into account.

Section 6 reviews the use of non-parametric statistics to assess pupil progress and school effectiveness in place of regression-based models of pupil value added. This includes particularly pupil value added computed from a comparison between a pupil's actual level of performance at GCSE or other Key Stage with the level of performance which would have been predicted for them on the basis of their prior attainment at the previous stage, such as at Key Stage 3 (KS3), using a 'median curve'. The median curve is mapped out by graphing the **national median level of performance** of pupils at the later stage, such as GCSE, amongst pupils nationally with similar prior attainments at the earlier stage, such as at KS3, against their prior attainment point score. School-level value added is computed by taking the arithmetic average of the pupil-level value-added measures. Studies by the DfES show that the extent of such pupil progress varies according to pupil gender, Free School Meals (FSM) eligibility, and ethnicity, as well as school-level variables, such as the proportion of pupils in the school eligible for FSM and the type of school involved. Systematically taking all of these additional influences into account, and assessing their statistical significance, in estimating the additional contribution which each individual school makes to pupil value added in the presence of variations across schools in these variables is difficult under a non-parametric approach, but remains feasible using a multilevel parametric model.

Section 7 reviews several studies of school value added which disaggregate the analysis to **individual subject areas**, and show both variations between schools in their levels of relative effectiveness across subjects, and also that variables such as pupil gender and pupil social background have a different influence on different subjects, such as English and mathematics. Other value-added studies have included not only school effects but also **teacher and class effects** when an extensive database on these intermediate variables has been available. In contract to the insights which these disaggregated studies reveal, aggregated single-level studies based upon regressing mean levels of school educational attainment on school mean levels of explanatory variables risk the 'ecological fallacy' of misinterpreting the resultant coefficients as confirming a significant relationship at an inappropriate level of the educational process.

Section 8 reviews the large number of studies in England and elsewhere which have examined the influence of **additional explanatory variables** in explaining pupil value added. These include not only pupil gender, but also pupil background variables that may reflect **socio-economic disadvantage** or initial difficulties when English is not the pupil's first language or a continuing influence of the pupil's junior school. In addition, they include **school context** variables which may reflect **peer group pressures** and wider social influences. The DfES's recent development of Contextual Value Added (CVA) models that incorporate many school- and pupil-level contextual variables using multilevel modelling, into the analysis of value added in different subjects at different stages of the educational process, represents a significant advance over earlier non-parametric methodology that did not take these variables systematically into account.

Section 8 also reviews other studies which have included **school resources** and **school processes**, and those which have examined the **stability over time** in annual estimates of individual school effectiveness. These have found positive, though imperfect, correlations over time in these annual estimates of school effectiveness, with different relative rates of individual **school improvement** or deterioration and some evidence of changes in the relative efficiency of different schools.

Different possible choices of the **functional form** for the value-added equation, and associated choice of **transformation of the variables**, are reviewed in Section 9. The merits of the logistic curve are discussed, alongside other choices such as a Cobb-Douglas formulation of

the educational production function, or a translog flexible functional form or a semi-log or logarithmic-reciprocal model. Rather than simply regarding the school effect disturbance term as an indicator of school effectiveness in a given single direction of educational attainment, **stochastic frontier analysis** seeks to distinguish such effectiveness from **heterogeneity** of the position of the underlying production frontier for each school due to additional unobserved factors. In contrast, **Data Envelopment Analysis** (DEA) adopts a non-stochastic approach to estimating a common production frontier for all schools that can take into account the multiple outputs which the school produces in different subjects and at different stages of the educational process. However, the **coefficient of technical efficiency** which DEA estimates for each school in this multiple output context does not itself provide detailed information on the effectiveness of the school in each relevant direction of educational attainment.

Section 10 examines extensive empirical studies which have been carried out into the possible existence of **differential slope parameters** on pupil prior attainment across different schools. Incorporating such a possibility through a random coefficients model allows greater consideration to be given to issues of **equality of treatment and of educational effectiveness across different ability groups** within the school. Mixed empirical evidence has been found in this context, though with differentiation by subject again showing interesting variations in school effectiveness. Other studies of differential school effectiveness have investigated the possibility of differential school effectiveness for different pupil groups differentiated by gender, ethnicity, FSM status or social class.

Section 11 examines several sources of possible **endogeneity bias** which may cause the parameter estimates of OLS or multilevel modelling to be biased away from their true values. The impact of several of these sources of endogeneity may be reduced by time lags and the concentration of published school league tables to date on absolute levels of school educational output levels, rather than chiefly value-added measures. However, one main remaining source of endogeneity arises if an intermediate measure of pupil attainment, such as at KS3, if used as the prior attainment measure for the explanation of pupil performance at a later stage, such as at GCSE, within the same school, with the level of school effectiveness within the error component analysis influencing both pupil performance at GCSE and at KS3. This suggests the need to avoid such a source of endogeneity by focussing instead upon each school's value-added performance between KS2 and GCSE.

Section 11 also examines the possible influence of **measurement errors** in biasing the estimated coefficients in a value-added analysis away from their true value in an underlying educational production function. However, knowledge of the true values in an underlying educational production function becomes less critical when value-added analysis is itself defined in terms of a comparison between achieved levels of pupil attainment and their **predicted** levels, conditional on the **observed** values of the explanatory variables. Nevertheless, there is a need for further research into the **sensitivity** of estimates of pupil- and school-level value added to possible variations in the observed data within the range of their likely inaccuracies, and the extent to which increasing the number of explanatory variables, rather than greater **parsimony** in their selection, reduces the **robustness** of the estimates to departures from the underlying assumptions of the model.

The developing literature on techniques of **programme evaluation** in conditions where **randomised control trials** are not feasible is examined in Section 12. In the absence of a random allocation of schools and pupils to the programme under a controlled experimental design, **selection bias** may arise that can bias the estimates of the programme impact that are generated by a technique such as multilevel modelling. The merits of other techniques, such as **difference-in-differences estimators**, are discussed, both in the context of a **homogeneous** (i.e. uniform) impact of the Academies programme on all schools participating in the programme and in the context of a **heterogeneous programme impact**. Section 12 also reviews the assumptions and implications of techniques of **matching**, such as the use of **Propensity Score Matching** under which schools would be matched that had the same probability of being selected for the Academies programme. In addition, the formulation of relevant **comparison groups** is discussed in Section 12. An important common feature of schools in the Academies programme is their low average level of **pupil prior attainment** at KS2, which, in line with value-added analysis, can be used as a key criterion to define relevant comparison groups.

A technique that has scope for application in the evaluation of the impact of the Academies when school- and pupil-level data are available both before and after the start of the programme is the use of a **regression-adjusted conditional difference-in-differences matching estimator**. This can generate consistent estimates of the programme impact under considerably weaker assumptions than those which are required under Propensity Score Matching. In addition, its use can be productively linked to a **value-added analysis** of school

effectiveness, both before and after the programme has been in operation, for both Academy schools, and their predecessor schools, and for schools in the comparison group. By adjusting measures of pupil attainment, such as examination results, for pupil prior attainment and other relevant pupil- and school-level variables, value-added analysis not only isolates more closely the contribution which the individual school makes to the pupil's educational progress, but at the same time corrects for many of the factors which would otherwise bias estimates of the impact which participation in an educational initiative, such as the Academies programme, has on those schools in the programme.

Several extensions to the application of relevant difference-in-differences techniques to the evaluation of the Academies programme are discussed in Section 13. These extensions include examining the impact of the programme on **disaggregated measures of examination performance** at different stages of the educational process and in different subjects, on **attendance and exclusions**, and on the **characteristics of their pupil intake**. In addition, they include examining the impact of the programme on **other secondary schools** and on their **Primary Feeder schools**.

Given the important contribution which value-added analysis can make both to programme evaluation and to the assessment of school effectiveness, there is a need for further research more widely into the impact which factors such as endogeneity bias, measurement error, choice of functional form, and parsimony in the selection of explanatory variables, can make to value-added estimates and their robustness, and into the relative merits of different estimation techniques in the face of these additional considerations. Whilst any conclusions based upon existing value-added models will be contingent upon the assumptions implicit in them, such further research can advance our existing state of knowledge of the effect of possible departures from these underlying assumptions.

1. INTRODUCTION

The evaluation of educational initiatives, such as the Academies programme, raises interesting questions as to the appropriate analytical tools and methodologies to be adopted in the evaluation. In this Research Report, we examine two main literatures which are relevant to this area. The first is the large and growing literature on the assessment of value added in the education sector. The second is the developing literature on programme evaluation in contexts where the conditions for carrying out randomised control trials (RCTs) are not fulfilled. We will examine both the opportunities and remaining problems that these approaches present for the evaluation of educational initiatives. In addition, we will seek to bring these two approaches productively together in providing appropriate analytical tools for the evaluation of educational initiatives, whether in a local, national or international context.

This review examines firstly the main approaches, conclusions, and issues of continued debate, associated with the current literature on the concept and measurement of value added in education, and especially in secondary education. Value added in its general economic sense refers to the extent to which the value of the inputs into the production process is increased when these inputs are transformed into the outputs of the production process. The concept of value added, or 'added value', is defined by Kay (1993) as "the difference between the (comprehensively accounted) value of a firm's output and the (comprehensively accounted) cost of the firm's inputs", arguing that "In this specific sense, adding value is both the proper motivation of corporate activity and the measure of its achievement". The computation of value added in this and related contexts, such as the assessment and administration of Value Added Tax, makes use of market prices to assess the value of inputs and outputs. However, many of the inputs and outputs of education have no simple market value. In the case of education, the inputs into the production process include not only resource inputs, but also pupils with different individual characteristics who do not have a direct market value.

While the concept of human capital (Becker, 1993) has been complemented by the computation of labour market rates of return on some stages of the educational process (see e.g. Dearden *et al*, 2000) and on studying some subjects, such as A-level mathematics (Dolton and Vignoles, 2002), significant problems remain for the computation of the economic value of each different level of educational attainment at each intermediate stages of the educational process before individuals enter the labour market (see Belfield, 2000). The relevance of

market values in assessing the value added by education may also be reduced by the traditional concern of education for equity of access and of provision for different levels of pupil ability, and for the development of capabilities other than those orientated towards the labour market.

2. THE VALUE ADDED CONCEPT

The broader economic issues are to some extent side-stepped by the more specific interpretation which has been given in recent years to the concept of value added in assessing performance within the education sector. This interpretation relates to the extent to which the pupils who are the subject of the value-added analysis are achieving the levels of educational performance that **might be predicted for them** at a given stage of the educational process, based upon information on their educational attainments at an earlier stage of the educational process, and upon other information that is considered relevant. For any given cohort C_{jg} of pupils in school j at stage g of the educational process, this gives a definition of their educational value added as:

$$v_{Cjg} = \sum_{i \varepsilon C_{jg}} [q_{ijg} - P(q_{ijg} \left| q_{ijg-r}, x_{ijg}, s_j \right.)] \qquad (2.1)$$

where q_{ijg} is a measure of the educational attainment of pupil i in school j at stage g of the educational process, $P(q_{ijg} | q_{ijg\text{-}r}, x_{ijg}, s_j)$ denotes the predicted value of q_{ijg} conditional on the **prior attainment** $q_{ijg\text{-}1}$ of pupil i in school j at a previous stage g-r of the educational process, x_{ijg} is a vector of other pupil characteristics that are considered to have an influence on the pupil's educational progress, and s_j is a vector of school-level variables that are considered relevant to the analysis.

A frequent means of deriving the predicted value, $P(q_{ijg} | q_{ijg\text{-}r}, x_{ijg}, s_j)$, is the use of **regression analysis**, in one of several different forms discussed below. The value added for each individual pupil is then the difference between their actual educational attainment score and that predicted for them, given $q_{ijg\text{-}r}$, x_{ijg} and s_j, by the regression line based upon a wider sample of pupils and/or schools. The definition of value added is therefore a **relative one**, of how well the pupils have progressed compared to what can be predicted or expected for them, given their prior attainment $q_{ijg\text{-}r}$, their other individual characteristics x_{ijg} and the school-level variables s_j, on the basis of a regression analysis of data on individual pupil achievements, q_{ijg} , and on $q_{ijg\text{-}r}$, x_{ijg} and s_j for pupils drawn from a wider sample of pupils and schools. The sum of the value-added measures for all the relevant pupils in the school yields the value-added score for the school as a whole in (2.1), when C_{jg} is taken to refer to all the pupils who have completed stage g of the educational process in the school j at the relevant date. In order to adjust for the size of

the school, the total value-added measure for the school may be divided by the number of pupils in the school who have completed stage g of the educational process in the school at the relevant date, to yield their average value added. A school will achieve a larger value-added score if the computed residual values of its individual pupil attainments at stage g, when compared to those that are predicted by the regression line, are greater. Such an outcome would be achieved, for example, by ensuring that their pupil achievements at stage g are all a larger distance above the predictive regression line. This basic approach is therefore referred to as **Residual Gain Analysis** by Fitz-Gibbon (1995).

The main roles which the value added concept is seeking to fulfil relate particularly to the assessment, and possible improvement, of **school effectiveness**. These potential roles include those of:

i. providing useful information to the management of each school on their performance relative to other schools in similar circumstances;

ii. providing information to parents on the school's educational effectiveness; and

iii. providing information to the wider public on the relative performance of the school to promote public accountability.

Much of the interest in school value-added measures in this context has arisen as a reaction to the perceived deficiencies of unadjusted school league tables that followed the requirement of the 1980 Education Act for secondary schools in England to publish their examination results. These presented summary measures of pupil achievements, q_{ijg}, within the school without making any allowances for **differences in pupil prior attainment levels** or in other variables within the vectors x_{ijg} and s_j that might be considered relevant in a value-added analysis. In contrast, a value-added approach to the assessment of school effectiveness implicitly regards the role of the school as adding value to pupils, who may start from different levels of prior attainment and have other relevant characteristics, that may influence their ability to achieve different levels of performance at the end of the given stage of the educational process.

The extent of the contribution of the school to the process of adding value can only then be gauged once the levels of prior attainment and the other relevant characteristics are taken into

account. Schools that serve pupils from disadvantaged backgrounds, but who none the less succeed in achieving examination results that are significantly higher than would be predicted for such a group of pupils, may then be judged as very effective under a value-added approach, but might be unfairly judged as ineffective if only their unadjusted position in school league tables was taken into account without regard to their disadvantaged pupil intake. In response to such criticism of unadjusted school examination results, the Dearing Report (1993) recommended the publication also of school value-added information as "a valuable contribution to appraising performance and to improving accountability".

A further important role for value-added analysis in the context of the Academies programme is:

iv. assessing the extent to which programme participation and changes in school status succeed in boosting the effectiveness of the schools concerned.

Because they are located in areas of social disadvantage, a value-added analysis approach has clear attractions in addressing the issues which are involved in this assessment. Since **iv.** in particular involves progress, and potential improvements in value added, *over time*, it is important that the value-added framework adopted can compare changes in value added over time in the Academies with the extent of the improvements which are taking place over the same period of time in appropriate **comparison groups** of schools**.** We will return in Section 12 to a more detailed discussion of the issues which are involved in specifying appropriate comparison groups, and associated techniques for programme evaluation. First, though, we will review the alternative approaches which have been adopted in the existing literature to the estimation of value-added measures.

3. MULTILEVEL APPROACHES

Performance data on pupils have a natural **hierarchical structure** in which pupils are located within classes within schools, that are in turn located within LEAs. Within this hierarchical structure, Aitken and Longford (1986) considered a number of formulations of the basic model for assessing school effectiveness under a value-added approach. The first was of the form:

$$q_{ij} = \alpha + \beta x_{1ij} + \gamma x_{2ij} + \varepsilon_{ij} \qquad \text{for } i = 1, ..., n_j \; ; \; j = 1,...,m \tag{3.1}$$

where q_{ij} denotes the level of attainment of pupil i in school j at the end of the relevant phase of education, such as a Key Stage, x_{1ij} denotes the level of prior attainment of the pupil at the start of the relevant phase of education, x_{1ij} is an additional pupil-level variable of interest, such as the gender or socio-economic background of the pupil, and ε_{ij} is a stochastic disturbance term that is assumed to be normally distributed with mean zero and a variance of σ^2. m is the number of schools in the analysis and n_j is the number of pupils in school j. α, β and γ are constant parameters to be estimated here using Ordinary Least Squares (OLS) regression analysis. An estimate of a '**school effect**' or '**coefficient of school value added**' can be obtained from the mean across pupils in the school of the residuals from an OLS regression, to obtain the predicted value of $q_{ij,}$, given the pupil's prior attainment level, x_{1ij} , and the value of the variable x_{2ij} , for each school j and pupil i. A model of the form of (3.1) was deployed in the U.S. Coleman (1996) report, which concluded that "schools are remarkably similar in the way they relate to the achievement of their pupils when the socioeconomic background of the students is taken into account ... it appears that differences between schools account for only a small fraction of differences in pupil achievement" (*ibid*, pp. 21-22).

One of the assumptions of the standard OLS regression model (see e.g Gujarati, 1995) applied to (3.1) is that the variance of the stochastic disturbance term ε_{ij} for each given value of x_{1ij} and x_{2ij} is the same constant, σ^2 . However, as Goldstein (1987, 1995) has stressed, the variance of distribution of ε_{ij} may well vary from school to school, with some schools facing more pupil-level variation than others. Such a variation in the pupil-level variance will introduce into (3.1) a element of **heteroscedasticity**. While they are still unbiased estimates if all relevant explanatory variables are included, the resulting OLS parameter estimates of the

regression coefficients α, β and γ will then not be **efficient**, i.e. **minimum variance** estimates in the class of linear unbiased estimators.

A second basic formulation of the value-added model of school effectiveness is that of:

$$q_{ij} = \alpha_j + \beta x_{1ij} + \gamma x_{2ij} + \varepsilon_{ij} \qquad \text{for } i = 1, \ldots, n_j \; ; \; j = 1,\ldots,m \tag{3.2}$$

where (3.2) replaces the constant intercept term α in (3.1) with a school-specific intercept term α_j to define a school **fixed effect** that represents the contribution of school i to pupil value added. (3.2) describes a set of parallel school-specific regression lines which differ from each other by variations in the extent of their school-specific intercept term α_j . The relative value-added by school j can now be computed as the value of α_j compared to the mean value of these school intercept terms across all schools. Due to the continued problem of heteroscedasticity, the parameter estimates for α_j , β and γ will still be subject to large **standard errors** if OLS is used for their estimation.

A third possible formulation which has been used in the literature on assessing school effectiveness, as in Marks, Cox and Pomian-Srzednicki (1983), is that of Aitken and Longford (1986)'s Model 3, involving the school-level **mean value** of pupil achievement, Q_j , for each school j together with the school-level **mean values**, X_{1j} and X_{2j} , of the pupil prior attainment levels and additional explanatory variable for school j, i.e.

$$Q_j = \alpha + \beta X_{1j} + \gamma X_{2j} + \eta_j \qquad \text{for } j = 1,\ldots,m \tag{3.3}$$

where η_j is a school-level disturbance term. If a form of weighted least squares is used to estimate the parameters of (3.3), the parameter estimates for β on pupil prior attainment in (3.1), (3.2) and (3.3) will in general differ according to the extent to which pupil-level variations in prior attainment level are due to **within-school** variations or **across-school** variations. This extent is itself likely to depend upon the degree to which **selection policies** operate in school admissions policies and/or the school attracts pupils from a **particular geographical area** that differs in its mean level of prior attainment from the geographical recruitment areas of other schools. If schools are relatively similar in their pupil intake, there will be less statistical variation in the school mean values on which to base reliable estimates of

the regression coefficients in (3.3), and on which to base the associated estimates of the school effect. On the other hand, if the within-school variation in pupil levels of prior attainment is small because pupils are recruited from segmented homogeneous geographical areas, the reliability of the estimate for β in (3.2) will be small.

Even aside from the problems which result from a high standard error to its parameter estimates, the school-level regression (3.3) may lead to the '**ecological fallacy**', of influences which operate differently at the pupil and school level being confused, when there are both school context variables and pupil-level influences operating on pupil attainment. At the same time, a '**means on means**' analysis of school level means, as in (3.3), may lead to a high correlation, and an apparently high variance explained, in a school level regression of educational performance on socio-economic background variables. However, if pupils are not randomly assigned to schools, but instead are selected or segregated by geographical areas with different socio-economic characteristics, both the regression coefficient and the estimated variance explained can be **biased** upwards by the means on means regression, compared to a multilevel analysis of disaggregated pupil data (see Fitz-Gibbon, 1996).

One approach to the inclusion of both **school context** variables and **pupil-level influences** is through the use of the school-level mean value, X_{1j} , of pupil prior attainment to capture the general level of pupil intake abilities within the school, alongside individual pupil-level prior attainments x_{1ij} in (3.1). When the variable x_{2ij} is omitted from the analysis, Aitken and Longford (1986) show that the coefficient estimate for β in (3.3) is simply equal to that on individual pupil-level prior attainment, x_{1ij} , in (3.2) plus that on the context variable X_{1j} in this extended version of (3.1). The estimates of the school effects in this extended model are, however, the same as those obtained from (3.3).

Rather than treating the school effects as fixed, when the number of schools in the sample becomes large enough to obtain reliable parameter estimates, it is of interest to express the school effects as linear functions of school-level variables, such as s_j , plus a school-level disturbance term, θ_j. We then obtain a formulation of the value-added model of the form:

$$q_{ij} = \alpha + \beta x_{1ij} + \gamma x_{2ij} + \delta s_j + \theta_j + \varepsilon_{ij} \qquad \text{for } i = 1, ..., n_j \; ; \; j = 1,...,m \qquad (3.4)$$

where θ_j is assumed to be normally distributed across schools, with zero mean and a positive variance of σ_θ^2 , but uncorrelated with ε_{ij} . The formulation (3.4) can be shown to introduce a positive covariance of σ_θ^2 between pupil attainment levels , q_{ij} , within the same school. The associated positive correlation between the residuals within the same school breaches a second assumption of the standard OLS regression model, namely that there is zero correlation between the disturbance terms ε_{ij} that correspond to different values of the explanatory variables. Another form of estimation process than OLS, such as Generalised Least Squares (GLS), is therefore required for the efficient estimation of the **variance component** model in (3.4). The estimation procedure of **Iterative Generalised Least Squares** (IGLS) which Goldstein (1987, 1995) advocates to implement this process may, however, in some circumstances fail to converge (McCullagh, 1989; Goldstein, 1987, 1995).

Raudenbush and Bryk (1989) consider the case where pupil performance depends differentially both on pupil-level background variable, such as prior attainment, and on its mean value, X_{1j} , within the school as a 'school context' or a school 'compositional effect'. Their model of pupil performance is then:

$$q_{ij} = \alpha + \beta_1 x_{1ij} + \beta_2 X_{1j} + \theta_j + \varepsilon_{ij} \qquad \text{for } i = 1, ..., n_j ; j = 1,...,m \tag{3.5}$$

where β_1 is the 'within-schools' regression coefficient and β_2 is the 'between schools' regression coefficient. Estimation of the single pupil-level model (3.1) or the fixed effect model (3.2) using data on x_{1ij} using OLS will both lead to estimates of the school effect that are biased even for large sample sizes, whenever β_1 differs from β_2 . While this asymptotic bias disappears for the two models when β_1 and β_2 are equal, or when the model (3.3) is fitted, use of OLS still produces **inefficient estimates** of the school effect in these cases.

Raudenbush (1989a) advocates use of pupil-level data that is **centred** around the school mean, such as X_{1j} , as in the formulation:

$$y_{ij} = \alpha + \beta_1 (x_{1ij} - X_{1j}) + \beta_2 X_{1j} + \theta_j + \varepsilon_{ij} \qquad \text{for } i = 1, ..., n_j ; j = 1,...,m \tag{3.6}$$

This can ensure that the pupil-level data are **orthogonal** to the corresponding school context variable, thereby amerliorating the problem of collinearity that may well otherwise arise between the variables x_{1ij} and X_{1j} in (3.5) and which would result in larger standard errors, and reduced precision, for the parameter estimates. A test for the importance of **school context** variables can then be obtained by testing whether the coefficient β_2 is **significantly different** from β_1 in (3.6).

4. THE MAGNITUDE OF THE SCHOOL EFFECT

In their empirical analysis of a sample of 907 pupils in 18 secondary schools (including two single-sex Grammar schools) within a single LEA, Aitken and Longford (1986), estimated an equation similar to (3.4), with pupil prior attainment in an ability test expressed as a Verbal Reasoning Quotient (VQR) as the single explanatory variable and pupil point score at CSE/O-level as the dependent variable, using a Maximum Likelihood estimation procedure. They found that the percentage of the overall variance in pupil performance that was accounted for by the school effect term was only **6.7 per cent** for the overall sample and only **1.9 per cent** for the more homogeneous sample of pupils in the 16 non–Grammar schools, with a correspondingly low value to the extent of the correlation between pupils' individual performance within the same school. The inclusion of other pupil- and school-level variables would have tended to reduce the percentage that is explained by the school-level random component even more, with the mean value of the pupil VRQ scores itself accounting for a large reduction in the estimated school effect when the pupil-level VRQ score is included in the regression.

Based on several subsets of a larger sample of approximately 14,000 individual pupils in 150 different schools within 6 LEAs covering urban, metropolitan and rural communities, Gray, Jesson and Sime (1995) used Iterative Generalised Least Squares (IGLS) to estimate equations similar to (3.4) that included pupil-level variables. Depending on which LEA was involved, the available pupil-level variables were subsets of pupil prior attainment on transfer to secondary school, pupil gender and measures of pupil background, including parental social class, housing tenure and the number of siblings in the family. Despite large differences in individual pupil performance in GCSE examinations, the variance that was attributed to (random) school effects before controlling for these pupil-level variables ranged **from 3.5 per cent in one LEA to 30 per cent in another**. However, when those pupil-level variables that were found to be significant were included in the fixed part of the equation, the percentage of the overall variance in pupil performance that was attributable to these fixed effects ranged from 10.8 per cent to 58.1 per cent for different LEA subsets, and that which was attributable to the (random) school effect ranged **from 1.5 per cent to 7.9 per cent** in nine of the eleven LEA datasets, **and 23.7 and 25.0 per cent** in the two datasets associated with LEA 2. LEA 2 was unusual in that it had retained informal selection for some of its schools, had a relatively

large selective voluntary sector and had very large social differences between some of the areas within the LEA, which were served by a small number of schools. These factors are likely to have increased the correlation between the performance of pupils within each school and the associated estimated school effect in the absence of pupil prior attainment data for LEA 2.

The relative low percentage of the variance in pupil performance that is attributed to the school effect in the other LEAs appears to be in line with the reported results of Nuttall *et al* (1989) of about **8 per cent** for secondary schools in Inner London and of Wilms (1987) for a wide range of Scottish secondary schools of no more than **10 per cent**. It is also in line with the findings of Thomas and Mortimore (1996), who found from a study of 79 secondary schools in Lancashire that "once background factors have been accounted for, the variation in pupils' total examination scores attributed to schools is 10 per cent" with corresponding figures of 9 and 12 per cent for English and mathematics respectively. However, this difference is not trivial. As Thomas and Mortimore (1996) note, "in terms of GCSE examination grades for individual pupils, this finding indicates an approximate difference of 14.4 GCSE points (that is, the difference between 7 Es and 7 Cs) between the most and least effective schools".

A generally small contribution of the school effect to the variance explained may also reflect a **low degree of variability** between schools in their teaching methods and organisation. As Montmarquette and Mahseredjian (1989, p. 190) note in their study of Montreal elementary schools: "The fact that class and school seem to have little effect on student achievement could be a result of insufficient variability in teacher and school input observed, and latent variables. It may be that in many schools systems, and, in particular in the case of Quebec, standardized teacher training, standardized teaching curriculum in the schools based on collective agreements, government rules and regulations concerning school organization and management, leave virtually no possibility for school effects to be detected by data analysis".

Raudenbush and Bryk (1989) argue that the estimates of the school effects will tend to be **biased downwards** if the school policy variables upon which school effectiveness is partially dependent are excluded from the value-added analysis, and that only if school policies are unrelated to variables which describe the composition of the pupil intake will estimates of these composition variables and the school effect be unbiased. They conclude that: "In principle, the direction of bias introduced by ignoring policy variables can favor schools which are either advantaged or disadvantaged on composition variables. However, experience suggest

that most often schools with advantaged student bodies will appear less effective than they are. Put another way, the relatively high achievement of advantaged schools will be attributed too much to the advantaged backgrounds of their students and too little to the effectiveness of the teachers and school policies".

A generally higher proportion (of around **15 per cent**) of variance that is attributed to schools was found by Fitz-Gibbon (1991) in a value-added analysis of A-level performance in chemistry, geography, French and mathematics, using pupils' average O-level grades as their prior attainment variable. The larger school effects at A-level than GCSE are attributed by Fitz-Gibbon to the potentially greater sensitivity of A-level grades to 'instructional effects'. They may also have been influenced by the absence of significant school context variables in the fixed component of the model for predicting A-level performance.

In a **three-way error component value-added analysis** of data from the U.S. National Educational Longitudinal Study that included **teacher, class and school effects**, Goldhaber *et al* (1999) concluded that "the vast majority of variance is explained by individual and family background characteristics (about 60%). Overall, school, teacher and class variables, both observed and unobserved, account for approximately 21% of the variation in student achievement. Of this 21%, only about 1 percentage point (or 4.8%) is explained by observable educational variables, and the remaining 20 percentage points (or 95.2%) is made up of unobservable school, teacher and class effects".

5. MULTIVARIATE APPROACHES

The relatively small proportionate influence of the school effect in several value added studies reported above also has relevance to the issue raised by Burstein in the published discussion that accompanied the paper by Aitken and Longford (1986), where he asserted that "the evidence is still out about how large the variance and covariance components must be practically to warrant choosing [the] more complete, complicated, and costly model and estimation procedure" involved in multilevel modelling.

The Final Report of the Value Added National Project (Fitz-Gibbon, 1997) stressed the desirability of adopting a simple and readily understandable approach to value-added assessment, based upon OLS regression models, for the purpose of providing internal school information on their relative performance. Recent estimates by Jesson (2001, 2002, 2003, 2004) and by Jesson and Crossley (2005, 2006) of the value added by Specialist Schools make use of the deviations of the individual schools' percentages of pupils attaining 5 or more A* - C passes from those that are predicted by an OLS regression analysis across all non-selective comprehensive and modern secondary schools. The OLS regression uses just two explanatory variables, the average KS2 point score of pupils five years before the date of their GCSE performance, and the proportion of boys in each school's GCSE cohort as a measure of the gender mix of the pupil cohort.

Feinstein and Symons (1999) have used OLS regression analysis to estimate a value-added model of pupil attainment at age 16 in secondary schools, based upon **pupil-level data** from the longitudinal National Child Development Survey of all children born in the UK between 3rd and 9th March 1958. Their model included pupil gender and prior attainment variables for reading and mathematics at age 11, as well as **family data** on father's socio-economic status and education and interest in the pupil's upbringing and education, the composition of the family, and on the mother's education and interest in the pupil's education. It also included **peer group** variables related to the proportion of children in the pupil's class with different characteristics, and school variables relating to the pupil-teacher ratio and school type. They found that "parenting is much more important than schooling. The most powerful parental input is parental interest in education .....We also find a strong peer group effect".

In a comparison between the estimates produced by using OLS and those which are produced under a multilevel error component estimations, using either Generalised Least Squares or Maximum Likelihood estimations, Montmarquette and Mahseredjian (1989) and Goldhaber *et al* (1999) find **little significant difference** between the estimates produced under such multilevel estimation, and those produced using OLS. Montmarquette and Mahseredjian (1989, p. 189) conclude from their study of pupil attainment in a sample of Montreal francophone public elementary school pupils that "The results are clear: for both test and grade levels the non-observable class variables are negligible in the explanation of school achievement. Latent school variables are more important, and even potentially interesting in first grade; but as long as student personal and socioeconomic latent variables account mostly for the residual component, these variables remain the best way to improve our understanding of student school achievement. This large residual component explains why generalized least-squares estimates do not differ from ordinary least-squares estimates".

Similarly Goldhaber *et al* (1999, p. 206) conclude that "the estimated coefficients from the random effects specifications of the models .... are very similar to those of the OLS specification .... In fact, there is only one case, that of teacher gender in the OLS model, in which a variable is statistically significant (at the 5% level) in one specification of the model and statistically insignificant (at the 5% level) in an alternative specification. There is also very little change in the magnitude of the estimated coefficients in the four models; thus, estimated returns to the schooling characteristics are relatively insensitive to whether the model is estimated with or without random effects, and insensitive to the specified level of the random effect".

This conclusion is in line with the expectation that the estimates of the fixed effects coefficients will be **unbiased**, both under OLS and error component estimation on **pupil-level data**. The **under-estimate of the standard error** of the coefficient estimates which OLS produces when multilevel random elements are important, compared to the use of error component models, may, however, result in some fixed effect variables being accepted as **statistically significant** under OLS which would not be found to be so under multilevel modelling. The extent of this under-estimation will tend to be greater, the greater is **the intra-school correlation** between the value-added residuals and the greater is the number of pupils per school. For intra-school correlations of 0.2 for both the pupil prior attainment and pupil achievement in a simple two-variable two-level model, with 76 pupils per school, Goldstein

(1995, p. 26) calculates that the standard error of the OLS estimate is a half that of the error component model. For a smaller number of pupils per school, Kreft (1996) reports the result of a **simulation study** based upon US SIMS data which **estimated the efficiency of OLS estimates at 90 per cent**, implying the need for more observations under OLS, and with **no differences** found for large samples between the efficiency under OLS and (restricted) IGLS estimates of the fixed effect parameters.

Multilevel models have the potential advantage that they also allow for explicit consideration of **differential random coefficients** across different schools on the underlying explanatory variables, in a way which cannot be incorporated into the standard OLS model. However, as we note below, the evidence for significant differential coefficients to date is limited. **Interaction terms** between different variables may also be studied within multilevel random coefficients (RC) models. However the simulation studies examined by Kreft (1996) indicated that for GLS, IGLS and OLS estimation methods equally large data sets are needed for these interaction effects to be detected, with "no proof ... yet found that RC modelling will help to discover interactions that could not be discovered with other methods".

Kreft (1996) also notes that the greater generality which multilevel models involve has its price. In particular under the iterative estimation procedures, such as IGLS that are used to estimate multilevel models, "larger data sets are needed to prevent instability of the solutions". In addition, "the estimation method used to estimate the parameter in the RC model is more complicated than in fixed effects linear regression models. Empirical Bayes maximum likelihood procedures are used to estimate the parameters of the model in an iterative process. Less is known of its properties".

An **iterative process** is required for the implementation of a Generalised Least Squares (GLS) approach under multilevel modelling because the block diagonal covariance matrix of the residual disturbance terms is unknown. The iteration process typically starts from the estimates of the fixed effect parameters based upon the multivariate OLS estimates, which assume zero correlations of the residuals within schools. Based upon these estimated fixed effects, estimates of the random components are obtained which enable a revised covariance matrix to be estimated and used to reassess the fixed effects using GLS procedures, which assuming normality will generate maximum likelihood estimates on convergence. However, particularly in small samples, these will produce **biased estimates** of the random parameters because no

account is taken in the procedure of the sampling variation of the fixed parameters (Goldstein, 1995). These estimates are therefore modified through the use of Restricted Maximum Likelihood (REML) within the Restricted Iterative GLS (RIGLS) procedure. Particularly where the explanatory variables enter in a non-linear way, **convergence** is, however, not always guaranteed (*ibid*, p. 79). **Instability of the parameter estimates** under this iteration process may then occur.

The estimates of the residuals which are produced by multilevel modelling are also adjusted for **sampling variation**, by applying a '**shrinkage factor**' to the unadjusted mean value-added residual for each school. This shrinkage factor (Goldstein, 1987, 1995) equals:

$$n_j\sigma_\theta^2(n_j\sigma_\theta^2+\sigma_e^2)^{-1}=1/(1+(\sigma_e^2/\sigma_\theta^2 n_j)) \qquad (5.1)$$

with the shrinkage factor becoming closer to unity as the number, n_j, of pupils in school j increases and as the ratio of the variance σ_θ^2 that is attributable to school effects to the variance σ_e^2 that is attributable to pupil-level variations increases. **Small schools** will then tend to have their **estimated school effect reduced** by this shrinkage factor, on the grounds that their results are more subject to sampling variation. The small size of the pupil sample in these schools, and associated larger sampling variation implies less confidence can be placed in an explanation that their results are due to a large individual school effect and a large estimated value added by the schools. No such shrinkage factor would be applied under an OLS multivariate regression analysis on pupil-level data that computed the school value added as simply the mean value within the school of pupils' individual value-added residual deviations from the estimated overall regression line. In contrast to the multilevel estimates of the value added by the school, the **OLS estimate** would not reduce a school's estimated value added downwards by such a shrinkage factor to allow for sampling variation and an assumed lower level of reliability of the value-added estimate for a smaller school.

In studies of value added for A-level students, Fitz-Gibbon (1991) estimates the shrinkage factor, that is used to indicate the reliability of the school effectiveness score prior to shrinkage, to equal 0.90 when the number of pupils sampled within a school is 50, 0.85 when it is 30, 0.65 when it is 10, and only 0.49 when it is 5. In comparing the estimates of the school value-added scores under multilevel modelling to those produced under OLS, she finds correlations

between the two sets of estimates ranging from 0.84 in the case of mathematics up to 0.97 in the case of geography.

The (shrunken) estimates of the residual value added scores which multilevel modelling produces under the above procedure are statistically **consistent**, i.e. approach their true value as the sample size increases towards infinity. However, for smaller sample sizes, they are **not unbiased estimates of the true underlying school value-added score** for any individual school (Goldstein, 1995, p. 24). As noted by Raudenbush in a personal communication to Fitz-Gibbon (1991):

> Although the shrinkage estimates are generally *more accurate* than estimates without shrinkage (their average distance from the true score is smaller than that of the non-shrinkage estimates), they are also *biased*. Suppose a school serving students with low prior attainment were especially effective. In this case, it would have its score "pulled" toward the expected value of schools with children having low prior achievement. That is, it would have its effectiveness score "pulled" *downward*, in the 'socially' expected direction, demonstrating a kind of statistical self-fulfilling prophesy! (On the other hand a similar school doing very badly would be pulled up.)

Kreft (1996) reviews simulation studies of how the multilevel modelling estimates of the variance component of the intercept that is due to higher-level (school) effects compare with the true values that are used to generate the dataset. She notes that "it can be concluded that RIGLS is less biased but also less efficient, while IGLS has more bias but is less efficient .... For both methods (IGLS and RIGLS) the variance components are underestimated or downward biased". A lower estimated variance of the value-added estimates for individual schools, and a lower associated assessed importance of the school effect, than their true values may therefore result under multilevel modelling.

6. NON-PARAMETRIC APPROACHES

A further approach to the assessment of value added is that published in the DfES Secondary School Performance Tables 2002 (DfES, 2003a), following the DfES's Pilot Value Added Study. This calculates value-added measures between KS2 and KS3, and between KS3 and GCSE/GNVQ, according to the following method. For the KS2 to KS3 measure of value added, the data used are those for pupils who were eligible for KS3 assessment in 2002 and on the school roll at the time of the KS3 assessment in 2002, and for whom matching KS2 prior attainment data was available. The value-added calculations exclude all pupils for whom results are disregarded at KS2 or KS3 according to Tables 6.1 and 6.2 below, with the exception that an input score of zero is used if the pupil was disapplied in all three subjects or had a combination of disregarded and disapplied results at KS2 and achieved at least one KS3 result at levels 2-7. The input and output measures that were used for each pupil in the value-added calculation were the numerical averages of the point scores which the pupil achieved in the English, Maths and Science results at KS2 for the input measure and KS3 for the output measure (or where any result is disregarded, the average of the remaining non-disregarded point scores at that Key Stage).

Key Stage 2 Level Outcome	Point Score: all subjects
6	39
5	33
4	27
3	21
Compensatory 2	15
N (not awarded a test level)	15
B (working below the level of the test)	15
Disapplied	Disregarded
Absent	Disregarded
Lost script	Disregarded

Source: DfES, 2003a

TABLE 6.1

Key Stage 3 Level Outcome	Point Score: English	Point Score: Maths & Science
E (for Exceptional performance)	57	57
8	51	51
7	45	45
6	39	39
5	33	33
4	27	27
3	21	21
Compensatory 2	-	15
N (not awarded a test level)	21	15
B (working below the level of the test)	21	15
Disapplied	Disregarded	Disregarded
Absent	Disregarded	Disregarded
Mixed Tier (maths & science only)	Disregarded	Disregarded
Lost script	Disregarded	Disregarded

Source: DfES, 2003a

TABLE 6.2

The value-added score of a pupil is then calculated by comparing the pupil's actual KS3 output measure with the **median level** of the KS3 output measure across the whole country of pupils in mainstream secondary schools with the same or similar input measure at KS2 to this pupil (with parallel separate calculations made for pupils in special schools). Table 6.3 below shows the median levels of the KS3 average point score output measure in mainstream secondary schools corresponding to different KS2 average point scores. Unlike multivariate regression analysis, where regression parameters would be used to predict the KS3 comparator for different average point scores at KS2, the approach here is the **non-parametric** one of selecting the corresponding **national median level of performance** at KS3 for each pupil-level average point score at KS2.

KS2 Average Point Score	KS3 National Median Average Point Score
0	21
15	21
17-18	21
19	23
21	27
23-24	29
25	31
27	35
29-30	37
31	39
33+	43

Source: DfES, 2003a

TABLE 6.3

The corresponding **median curve** of the KS3 pupil-level National Median Average Point Score (NMAPS) against the KS2 Average Point Score is not a straight line, being in particular flat at a value of 21, such as that achieved by a Level 3 outcome at KS3, for KS2 Average Point Scores between 0 and 18. The value-added score of a school between KS2 and KS3 is calculated as 100 plus the arithmetic mean of the value-added scores of all the pupils in the school for whom the value-added scores are calculated, rounded to one decimal point.

A similar approach has been used to calculate value-added scores between KS3 and GCSE/GNVQ. The point scores at GCSE is derived from the GCSE grade according to Table 6.4, with an equivalence table used to calculate corresponding point scores from GNVQs. Each pupil's GCSE/GNVQ output measure is calculated by **summing their best 8** GCSE/GNVQ point scores, and disregarding any other of their less good GCSE/GNVQ scores.

GCSE Grade	GCSE Points	GCSE (Short course) Points
A*	8	4
A	7	3.5
B	6	3
C	5	2.5
D	4	2
E	3	1.5
F	2	1
G	1	0.5
U,X	0	0

Source: DfES, 2003a

TABLE 6.4

The input measure at KS3 for each pupil whose KS3 record can be matched with a corresponding GCSE/GNVQ score is calculated as the numerical average of their KS3 point score achieved in the English, mathematics and science tests, and the output measure. Their **value-added score** between KS3 and GCSE/GNVQ is the difference between their GCSE/GNVQ output measure and the **national median level** of the GCSE/GNVQ output measure of all pupils with the same or similar input measures at KS3. The national median 'best 8' GCSE/GNVQ output scores for pupils in mainstream schools are shown in Table 6.5 below. Each school's value added is again 100 plus the arithmetic mean of the value-added measures that have been calculated for pupils in the school, rounded to the nearest decimal point.

KS3 Average Point Score	National Median 'best 8' GCSE/GNVQ point scores
0-18	6
19	9
21	11
23-24	16
25	20
27	24.5
29-30	29
31	33
33	37
35-36	40.5
37	44
39	47
41-42	51
43	54
45	58
47-48	60
49	62
51+	63

Source: DfES, 2003a

TABLE 6.5

It should be noted that the above national median output levels are taken across **all pupils, both girls and boys, and irrespective of the proportion of pupils eligible for Free School Meals (FSM) or other variables related to the characteristics of the pupil intake**. However, several studies by the DfES (2001, 2002, 2003b) show that such **additional variables** may well be relevant to the **statistical explanation** of the wide variations that exist in pupil attainment at each secondary Key Stage between pupils with similar prior attainments

at the previous Key Stage. DfES (2003b) finds that at GCSE **girls** make more progress than boys for each KS3 level of prior attainment, with a difference of 15 percentage points in 2002 in the percentage of pupils attaining 5 or more A*-C grades at GCSE for pupils with a level 5 at KS3, and with girls making more progress in English than boys at all Key Stages. Similar significant differences in rates of pupil progress by **gender** were found in the earlier DfES (2001) study based upon end-of-Key-Stage assessments in 2000.

The DfES (2003b) study also finds that **non-FSM** pupils make more progress from each prior attainment level in each subject at every Key Stage than do FSM pupils. Moreover, while pupils for whom **English is an additional language** (EAL) start below the national median level, they in general make more progress than non-EAL pupils, tending to overcome some of their initial relative disadvantage as they become more proficient in English. Once they are performing above the national median level, they tend to make progress in line with non-EAL pupils. Rates of progress also tend to vary with **ethnicity**. Amongst boys, Chinese pupils progress most at all Key Stages, while amongst girls, Pakistani pupils start as one of the least progressing groups at KS2, but at GCSE are one of the best progressing groups. Black Caribbean pupils, whether boys or girls, or FSM or non-FSM pupils, are found to make below average progress at all Key Stages. White pupils are found to be one of the worst progressing ethnic groups at GCSE, and have the greatest difference (of 16 percentage points) between FSM and non-FSM pupils in progress to KS3 English and Science from the national median KS2 level.

School-level variables also appear to influence the rate of pupil progress. Non-FSM pupils are found by DfES(2003b) to make more progress in schools with a lower overall proportion of pupils with FSM, as do FSM pupils. In every subject at KS3, FSM pupils in low FSM schools make more progress than non-FSM pupils in high FSM schools, suggesting that either a strong peer-group effect may be present or that non-FSM pupils in high FSM schools may still suffer from deprived background variables that are imperfectly reflected in the FSM indicator. A similar tendency is found in English, science and reading at KS2, but not at GCSE. DfES (2002) similarly found that pupils in low FSM schools made greater progress than pupils in high FSM schools, especially between KS2 and KS3 where there was found to be up to half a level difference between the median progress in the highest and lowest FSM bands of schools.

Other **school-type** factors were found by DfES (2002) to make less difference than the proportion of pupils eligible for FSM, and appear to account for only a small proportion of the wide spread of outcomes for pupils with similar prior attainments. Between KS2 and KS3, pupils in Voluntary Aided schools tended to make above average progress in English, while pupils in Foundation schools tended to make above average progress in mathematics, and pupils in Voluntary Controlled schools tended to make above average progress in science. Between KS3 and GCSE, pupils in Voluntary Aided schools and in Foundation Schools made above average progress, while those in Voluntary Controlled schools made below average progress in general. DfES (2002) also found that on average pupils in secondary faith schools made similar progress to pupils in all secondary maintained mainstream schools, except for slightly superior progress in KS3 English.

Between KS2 and KS3, in each core subject, pupils in **specialist schools** on average made slightly more progress than pupils in all maintained mainstream schools. During KS4 they made 1 or 2 GCSE points more progress than pupils in all schools. Pupils in **Beacon Schools** made a fifth of a level more progress in English and a tenth of a level more progress in mathematics and science between KS2 and KS3 than pupils in all secondary maintained mainstream schools. At GCSE/GNVQ, pupils in Beacon Schools at the lower end of the prior attainment range made on average 4 GCSE points more progress than in all maintained mainstream schools, whereas at the upper end of the prior attainment range DfES (2002) found that they made on average 2 GCSE points difference. In mathematics and science, pupils in schools in **Education Action Zones** (EAZs) were found to make similar progress between KS2 and KS3 to pupils in all secondary maintained mainstream schools with broadly equivalent FSM proportions, whereas in KS3 English they made on average an eighth of a level less. Between KS3 and GCSE, pupils in EAZ schools also made similar progress to pupils in all secondary maintained mainstream schools with broadly equivalent FSM proportions, except that at the highest levels of prior attainment, pupils in EAZ schools made on average about 1 GCSE point more progress and those at the lowest levels of prior attainment, pupils in EAZ schools made on average about 1 GCSE point less progress. In addition for pupils in schools with FSM percentages in the range 21-35 per cent, EAZ schools made about 1 GCSE more progress.

Pupils in schools in **Excellence in Cities** (EiC) schools were found by DfES (2002) to make more progress in KS3 English than pupils in all secondary maintained mainstream schools with

broadly equivalent FSM proportions, with a difference that increased to about an eighth of a level at the highest levels of prior attainment. A similar difference at the upper levels of prior attainment was found in KS3 mathematics and science, but pupils with prior attainments of below an average level 5 in EiC schools made similar progress in KS3 mathematics and science. However, between KS3 and GCSE, pupils in EiC schools made about 1 GCSE point more progress for similar prior attainment levels than pupils in all secondary maintained mainstream schools with broadly equivalent FSM proportions, but with this difference increasing to about 2 GCSE points at the extreme ends of the prior attainment range.

Pupils in non-selective **single sex** schools were found to make on average more progress than pupils of the same gender in non-selective **mixed** schools at both KS3 and KS4, particularly in English. Between KS2 and KS3, pupils in designated selective LEAs who were at the upper end of the prior attainment range on average were found to make a quarter of a level more progress than pupils elsewhere, but similar progress in core subjects at the lower end of the prior attainment range. However, between KS3 and GCSE/GNVQ, pupils throughout the prior attainment range made about 1-2 GCSE points less progress in the designated selective LEAs than elsewhere. **School size** also appears to have some impact on pupil progress. Pupils in schools with year group cohorts of less than 100 pupils made less progress, even after adjusting for differences in FSM proportions. However, the size of the year group cohort appears not to affect the rate of pupil progress so long as it is greater than 100.

Rates of pupil progress may also **vary over time**. Between 2001 and 2002, DfES (2003b) found that the rate of pupil progress at KS3 decreased in each subject measured in terms of the percentage of pupils attaining either KS3 level 5 or above, or level 6 or above, for every KS2 prior attainment level. The reason given for this is that the improvements made at KS2 between 1998 and 1999 have not fed through to similar improvements at KS3 from 2001 to 2002. Progress at KS3 has also decreased in English and mathematics for almost all KS2 prior attainment levels between 1999 and 2002. A decrease in the rate of pupil progress is also found between 2001 and 2002 of 3 and 2 percentage points in those gaining 5 or more GCSE A*-C grades for KS3 prior attainment levels 5 and 6 respectively, although progress has improved by 2 percentage points for KS3 prior attainment levels 4 and 5 between 1999 and 2002.

The above studies in DfES (2001, 2002, 2003b) suggest that a number of pupil- and school-level variables, such as gender, FSM, ethnicity and school size, may be influencing the rate of

progress of individual pupils with similar levels of prior attainment. If this is so, there is a weaker case in assessing school effectiveness for computing school value-added measures, as in DfES (2003a), as the sum of the pupil-level differences between their actual performance at a given Key Stage and the **national median level of performance** for the national cohort of students with a similar level of prior attainment, **irrespective of these other pupil- and school-level variables**. If these value-added measures are to be used as measures of school effectiveness, or changes in school effectiveness following Academy status, there is a strong case for taking into account those additional variables which have a **significant systematic influence** on rates of pupil progress for the same level of prior attainment.

One, non-parametric, way of doing this would be by comparing each pupil's actual level of performance at a given Key Stage with the median level of performance of those pupils nationally with **similar levels of prior attainment and similar levels of these other statistically significant variables**, including potentially those corresponding to gender, FSM at pupil and school level, ethnicity, and school-type. However, if the number of such additional relevant variables is large, there may be a relatively small number of pupils in each relevant cell, with which to compare each pupil's performance.

The above studies, moreover, do not provide any formal statistical tests of the significance of different pupil and school level variables in influencing the rates of pupil progress. In contrast to multivariate regression analysis or multilevel modelling, the **piecewise comparisons** which they involve, of the apparent influence of different variables on median levels of pupil progress, are less well suited to the identification of the relative strength and significance of each of several different pupil and school level variables which may be acting simultaneously on the rate of pupil progress in different subjects. It should also be noted that the KS3 and GCSE/GNVQ point scores used in the above analyses were based upon the levels that pupils achieved in their Key Stage assessments and progress in **discrete steps**. Pupils who achieved level 4, for example, received 27 points, and those who attained level 5 received 33 points. Small changes in the underlying distribution of pupil marks can then result in apparently large, and seemingly significant, changes in average point scores and the associated **median values** of the average point scores. A mean value based upon the underlying marks would instead be more stable and incorporate more detailed quantitative information than the median value based upon average point scores based upon more discrete levels. As an indicator of the effect of different pupil- and school-level variables, the median value is also less sensitive to pupil

performance at either end of the range than the mean value, in circumstances where arguably both **high and low levels** of educational performance should be fully taken into account in assessing the overall impact of different variables.

Although the above non-parametric studies avoid explicit modelling of the multivariate influences of the different variables on pupil progress, the problem of **endogeneity bias** does not disappear as a result of a lack of such explicit modelling. As we note below, one important possible source of endogeneity bias is the existence of a correlation between pupil prior attainment at KS3 and the school-level disturbance term that may reflect school effectiveness not only at GCSE but also at KS3. Studies of value added from KS3 to GCSE that draw heavily on data where KS3 and GCSE examinations are both taken within the same school may then be particularly subject to such endogeneity bias.

Since pupil progress may indeed depend upon additional pupil- and school-level variables, a more explicit multivariate parametric estimation procedure has more recently been pursued by the DfES as part of its **Contextual Value Added** (CVA) project. This is discussed more fully in Section 8 below.

7. AGGREGATION ISSUES

The desirability of **disaggregating** data to the lowest level at which the variables are likely to have their impact is emphasised by the misleadingly high correlations which can occur between aggregated data at a higher level of analysis. As our earlier discussion of the '**ecological fallacy**' emphasised, regression analyses of **school-level mean values** of educational performance on school-level mean variables, such as school context variables, may yield biased estimates both of the regression coefficients and the variance explained, compared to disaggregated estimates based upon pupil-level data (see also Woodhouse and Goldstein, 1988; Woodhouse, 1990; Fitz-Gibbon, 1996). The problem of **aggregation bias** when the coefficients in individual micro-relationships differ across individuals was emphasised by Theil (1954). While Zeller (1966) showed that in a random coefficient model an OLS estimate of the aggregate relationship can provide an unbiased estimate of the common mean of the coefficient, inferences based upon the associated standard error of the estimate are likely to be misleading. Moreover a non-random allocation of individuals to schools, and the existence of school effects that are correlated with the individual explanatory variables, will add further sources of bias in the parameter estimation in the context of value-added analysis.

The desirability of disaggregating educational outcome measures into measures of performance in **different subject areas** is emphasised by a number of empirical studies of school value added. Several existing studies (e.g. Trower and Vincent, 1995; Fitz-Gibbon, 1997) provide evidence that schools' relative performance, particularly at secondary school level, varies significantly across different subjects. Moreover, some variables, such as **gender**, may be particularly important as explanatory variables in explaining pupil achievement in some subjects, but not others. Thus Smith and Tomlinson (1989) found no significant gender difference in overall examinations scores, but found that girls performed significantly better in English for each level of prior attainment than did boys.

Smith and Tomlinson's (1989, p. 282) study of *The School Effect* concluded that "School differences are greater when the results in particular subjects are considered than when the results are considered in total across all subjects. There is a considerable tendency for the schools that do well in English to do well also in maths and across all subjects, but some schools do far better in one subject than the other, or in one subject than across all subjects. All of these findings can be explained if it is true that the style and content of teaching is

determined more at the departmental level than at the school level". Based upon their variance-component analysis of twenty multi-racial comprehensive schools during the 1980s, they concluded that "the analysis shows large differences between the exam results achieved by children at different schools. A statistical model predicts that a boy with an above-average second year reading score of 110 would just fail to get a CSE grade 3 in English if he went to school 14, but would get an O level grade B in English if he went to school 25. Differences between schools in their maths results are of similar size. There are also large differences between schools in terms of the exam results they achieve in total across all schools" (*ibid*, pp. 281-2). Trower and Vincent (1995) similarly concluded from their empirical study of pupil value added by GCSE results in 1996 matched to 1994 KS3 prior attainment that "A value-added indicator based on an outcome measure which combined all subjects would hide the **substantial differences in success** between different departments within the school". Thomas and Mortimore (1996) similarly found "strong evidence that schools are not consistently effective in the two subject areas analyzed, mathematics and English", with a correlation between individual schools' value-added scores in English and in mathematics, within the single LEA studied, of only 0.46.

In addition, van de Werfhorst *et al* (2003) have emphasised the scope for a differential impact of pupil **social background variables** across different subjects, arguing that "students from homes lacking in 'cultural capital' may find it harder to compete in arts and humanities subjects than in scientific and technical subjects, where they do not face the same comparative disadvantage. The effects of home background may be comparatively important for arts and humanities subjects, whereas school effects have more of an impact on attainment in mathematics and sciences", citing the findings of Shaycroft (1967), Coleman (1975), Postlethwaite (1975), Brimer *et al* (1977), Mortimore *et al* (1988), and Brandsma and Knuver (1989) in support of this hypothesis. Their hypothesis is also supported by the findings of Sammons *et al* (1993) in a study of London junior schools that "background factors and initial attainment account for more of the variance in reading than in mathematics attainment and that school differences may be greater for mathematics". The desirability of disaggregating the value-added analysis by subject in such circumstances is underlined by the relatively low correlation, of 0.62, which Sammons *et al* (1993) found between the estimated school effects in reading and mathematics, with only four schools out of the sample of 49 having significantly positive residuals for both subjects, though with six schools having strongly negative residuals for both.

The issue of the appropriate level of aggregation or disaggregation of the data also arises in the context of the appropriate levels and **units of analysis** in a multi-level context. Whilst many multilevel studies, such as that by Smith and Tomlinson (1989), have focussed on the identification of the random component of the school effect within **two-level** model of pupil- and school-level variables, Montmarquette and Mahseredjian (1989) consider also the intermediate level of *classes*, and identify separate school and *class effects*, using a sample of francophone Montreal public elementary school pupils. As noted above, they found the percentage of the variance in pupil achievement that are explained by a random class-level component to be small, and that by the random school-level component to be 6 per cent or less. However, unobserved personal and socio-economic components, rather than observed fixed effects, explain from 60 to 75 per cent of the variance in pupil achievement. In a review of several other studies of school- and class-level effects, Hill and Rowe (1996) conclude that their results "suggest that variation between classes is far more significant than variation between schools, although in detail the evidence often appears to be contradictory and open to a variety of interpretations", such as a failure to adequately adjust for streaming pupils of different abilities into different classes.

Individual **teacher and school effects** were also estimated using an error component model in the Tennessee Value-Added Assessment System (TVAAS) (Tucker and Stronge, 2005), using an extensive database of pupil and teacher records, with attendance records used to provide details of whether a pupil has been present in each teacher's class for at least 150 days in the year in order to link individual pupil progress to individual teachers in different grades and different subjects (Sanders and Horn, 1994). Sanders and Horn (1998) conclude that: "Research conducted utilizing data from the TVAAS database has shown that race, socioeconomic level, class size, and classroom heterogeneity are poor predictors of student academic growth. Rather, the effectiveness of the teacher is the major determinant of student academic progress. Teacher effects on student achievement have been found to be both additive and cumulative with little evidence that subsequent effective teachers can offset the ineffective ones".

A study which carried out a three-way error component analysis, through decomposing the variance in pupil performance into school, class and teacher effects, is that by Goldhaber, Brewer and Anderson (1999). One of the valuable features of the U.S. National Educational

Longitudinal Study (NELS) data on which this study is based is that it includes detailed teacher- and class-level data linked to individual pupils, including each teacher's gender, ethnicity, experience, degree level and certification, as well as class size and composition. Their value-added analysis was based upon the 10th grade mathematics scores of a random sample of pupils drawn from 490 different schools, involving 1340 classes and 1089 different teachers, using pupil-level prior attainment scores in mathematics at Grade 8, as well as other pupil-, class-, teacher-, and school-level variables discussed in Section 8 below. They concluded that "The vast majority of variance is explained by individual and family background characteristics (about 60%) Overall, school, teacher and class variables, both observable and unobserved, account for approximately 21% of the variation in student achievement. Of this 21%, only about 1 percentage point ... is explained by observable educational variables, and the remaining 20 percentage points ... is made up of unobservable school, teacher and class effects" (of about 8, 8 and 4 per cent respectively). In a similar study using NELS, Goldhaber and Brewer (1998) examined the possible **omitted variables bias** that arises from the exclusion of typically unobservable teacher effects, by making use of information that is available within NELS on teacher behaviour, such as the percentage of time devoted to small group teaching, to maintaining order and to administration, and to their teaching style, preparedness and disciplinary policy. They find that these variables do not appear to be correlated with more generally available teacher characteristics, such as years of experience, and hence do not bias the estimates of the impact of these teacher characteristics on pupil attainment when the teacher behavioural variables are omitted.

Using aggregate data at the level of U.S. states, Card and Krueger (1992a and 1992b) concluded that the wage premium associated with an additional year of schooling tended to be associated with smaller classes and higher teacher salaries. However, Hanushek, Rivkin and Taylor (1996) criticised this conclusion on the grounds that the use of such aggregate data resulted in greater model misspecification. They concluded that "The results in [our] paper provide evidence that problems of omitted variables bias tend to increase along with the level of aggregation, causing analyses that use more aggregate data to overestimate systematically the influence of school expenditure related characteristics on student attainment. Aggregate analyses of student performance, particularly at the state level, typically have very crude measures of school and family factors. They never employ value-added models. Moreover, aggregate analyses drawing data from different states generally neglect potentially important

financing, organizational, and regulatory features of states. In short, they are subject to excessive specification problems".

8. THE EXPLANATORY VARIABLES

a. Pupil background

The set of explanatory variables that have been used in empirical studies of school effectiveness have included not only measures of prior attainment (such as test results at an earlier Key Stage, or earlier reading, verbal reasoning and mathematics test scores) but also variables related to pupil background (such as measures related to parental income and education, ethnicity and gender). In those US studies, such as Coleman, Hoffer and Kilmore (1982), which have used only pupil background variables without the use of pupil prior attainment data, Gray (1989) concluded that "controls for differences in pupils' backgrounds appear to 'explain' between 20-30 per cent of the variation in pupils' outcomes .... In contrast, prior attainment studies typically 'explain' upwards of 50 per cent of the variation in pupil outcomes .... In the ILEA's Junior School Project, prior attainment explained as much as 61.7 per cent of the variation in reading attainment three years later. Knowledge of background factors (notably social class and ethnic background but also fluency in English and free school meals, this last a measure of economic disadvantage) add just 2.3 per cent. Another British study shows background measures (father's occupation and family size) adding just three per cent to variance explained, once verbal reasoning has been taken into account (Daly, 1986); while Rutter's [1983] evidence indicates that father's occupation alone adds a mere 0.1 per cent".

The comparisons made in two studies by Willms (1986, 1987) of the impact of including either prior attainment measures or pupil background variables or both as explanatory variables for pupil outcomes indicate a large bias in the estimated school effects if only pupil background variables are included, compared to pupil prior attainment also being included as an explanatory variable. However, when pupil background variables are excluded and only prior attainment is included in the explanatory variables, the degree of bias found in the estimated school effects, compared to when all of these variables are included, is only a small one.

In the view of Gray (1989), the apparently large 'explanatory gap' of 20-30 per cent which appears available for explanation if only pupil background variables are included, though not if prior attainment measures are included, has "contributed to researchers overstating the extent

to which they need to be interested in complex estimation procedures in which varying slopes and interactions by slopes are considered. I suggest we proceed slowly in declaring such interests – they introduce a degree of complexity that may deter potential users".

As noted in Section 3 above, one set of potential candidates for inclusion is that of 'school context' variables, such as the school mean levels of pupil prior attainment and of other pupil-level variables. These variables may reflect **peer group pressures** on individual pupil performance. Rutter *et al* (1979) concludes that schools with higher proportions of more able pupils appeared to achieve better levels of individual performance for pupils with similar prior attainment scores than schools where the mean level of performance was lower. Similarly, Coleman *et al* (1966) concluded that individual pupils with similar levels of **social disadvantage** performed better in schools where the mean level of these background variables was more favourable. However, these findings have been criticised, for instance by Jencks *et al* (1972), Smith (1972) and Heath and Clifford (1981), as being potentially due to inadequately measured pupil-level variables, rather than necessarily to peer group effects. Although the inclusion of such context variables was found to substantially reduce the estimated residual school effect for several LEAs, Gray, Jesson and Sime (1995) found little evidence of these context variables being significant for those LEAs which had extensive pupil-level prior attainment data that could be included within the analysis.

Sammons (1995) found that, in contrast to their progress in primary school reading and mathematics, membership of one of the three **ethnic minority groups**, of Asian, Carribbean or all other backgrounds, was associated with significantly greater value added during secondary education up to GCSE than was membership of the majority ethnic group of English, Welsh, Scottish or Irish background. While at age 10 an Asian or Carribbean background was found to be associated with a significantly lower reading score, and a Carribean (though not an Asian) background with a significantly lower mathematics score, during secondary education members of these ethnic minority groups tended to make greater relative progress in terms of value added than the ethnic majority group.

Thomas and Mortimore (1996) compared the results of applying several different value-added models of secondary school effectiveness to GCSE data from 79 schools in one LEA. Their first model made use only of pupil-level data. In explaining the total performance score and the English score, the significant pupil-level variables included gender, ethnicity, pupil age, NFER

prior attainment verbal, quantitative and non-verbal tests on entry to the school, FSM status, the number of years in UK secondary schools and whether they had attended more than one secondary school. In explaining the mathematics score, the pupil gender proved insignificant. The addition of a quadratic term in the quantitative prior attainment score for total examination performance and in the verbal prior attainment score for the English results also improved the goodness-of-fit of the model.

A second version of the model involved examining the impact of 30 additional variables drawn from the 1991 census, matched firstly to individual pupils on the basis of each pupil's home postcode and secondly matched to the postcode of each school ward. None of the school ward census variables was found to be significant when the pupil-level census variables were included at the same time. Only two pupil-level census variables were found to be significant after non-significant census variables were eliminated, namely those relating to the percentage of households in unskilled occupations and the percentage of households with higher qualifications. When aggregated pupil-level data, such as the percentage of pupils entitled to FSM, were also included, they were not found to be significant alongside these two pupil-level census variables.

A third version of their model made use only of **pupil prior attainment data** and this alone was found to reduce the estimated variation in school effectiveness by **70.6 per cent** compared to if none of these variables was included. The additional pupil-level variables that were included in their first model boosted this to **72 per cent**, and the pupil-level census variables that were also included in the second version of the model boosted it to **75.2 per cent**. When only prior attainment data were included in the analysis of pupil performance in mathematics and English, the percentage reduction was greatest for mathematics, indicating that pupil gender, ethnicity and socio-economic factors are more important for GCSE performance in English than in mathematics.

The fourth version of their model excluded data on pupil prior attainment, and now found five pupil-level census variables and one school context measure (the % of pupils entitled to FSM) to be significant. It also resulted in changes in the significance level of other pupil-level variables, such as ethnicity, with the assessed negative impacts of the pupil-level FSM variable and the variable indicating more than one secondary school attended increased, and the positive impact of ethnicity for most non-white pupils (except Chinese) substantially reduced,

by failing to include pupil-level prior attainment. From their analysis, Thomas and Mortimore (1996) conclude that "the context of the overall school has little impact on the variation in pupil outcomes once adequate account has been taken of pupil intake factors (outside the control of the school).....However, when pupil intake data are lacking, school context factors (such as % FSM) may provide useful and adequate proxy measures of the level of attainment of pupils on entry to school".

An additional pupil-level factor which Goldstein and Sammons (1997) found to be an important predictor of GCSE performance, in addition to pupil prior attainment, was the **junior school** which the pupil attended (see also Sammons, 1999). Goldstein (2001) notes that "a pervasive problem for all analyses of performance is pupil mobility. Pupils who change schools tend to have lower test scores on average. If a school has a high proportion of such children, then raw scores will tend to lower perceived performance and value added scores may have an upward bias if computed solely from those pupils present at the start and end of the particular stage of schooling".

Goldhaber, Brewer and Anderson (1999) used **pupil-level** data on prior attainment in mathematics at grade 8, gender, ethnicity, family income, parental education, and whether the pupil was living only with their mother or their father. In addition, they used not only **school-level** data on whether the school was located in a rural or urban area, its size, the percentage of pupils who were white, the percentage of pupils who were from single-parent families, and the percentage of teachers with a Masters degree or higher, but also **teacher-level** variables on the gender of the teacher, the ethnic background of the teacher, their years experience in secondary education, whether they were a certified teacher, and whether they had a Masters degree or higher, together with **class-level** variables on class size and the percentage of ethnic minority pupils in the class. The variables which were found to be significant at the 5 per cent level in influencing pupil performance in mathematics at Grade 10 were the prior attainment score, parental education, and class size, each with a positive coefficient, and the percentage of pupils in the school who were white and the percentage of ethnic minority pupils in the class, each with a negative sign. The positive sign to class size, however, suggests that the potential endogeneity of class size may not have been adequately corrected for in this analysis.

b. Gender

Gender differences in educational progress have been found in both primary and secondary schools. In a follow-up study to the *School Matters* project of Mortimore *et al* (1988), Sammons (1995) carried out a multilevel value-added analysis of longitudinal data on the educational performance and characteristics of age-cohort of pupils over a nine year period from their year 3 performance in primary school through to their GCSE performance in year 11. At age 7, she found significant **gender differences in favour of girls** in reading performance but not in mathematics. At age 10, boys were, however, performing significantly less well in mathematics by age 10, with **gender an important predictor of value added** in mathematics, though not in reading, during these years. Boys also showed significantly lower value added in general compared to girls during secondary schooling up to GCSE, so that overall disparities in absolute levels of performance between boys and girls increased during secondary education up to age 16.

c. School resources

Based on a single cohort of 29,544 students in 690 schools in the fourth grade in Alabama in 1990-91, Ferguson and Ladd (1996) estimated a multilevel (or 'linear hierarchical') value-added model for standardised pupil test scores in reading and mathematics. Their dependent variables included the individual pupils' prior attainment scores in the same subjects in grade three, pupil-level gender, age and ethnic variables, the average fourth grade class size in the school, the average characteristics of fourth grade teachers in terms of the percentage with more than five years' experience, their own test scores and the percentage with a Masters degree. In addition, they made use of school data from school administrative records and school district data from 1990 census data based on the local zip code level. The school and district data included data on parental education, family income, the percentage of pupils in the school who were eligible for free or reduced-price lunch, and the percentage of pupils in the school's fourth grade who were not in the same school in the third grade as a measure of pupil mobility, together with data on school district enrolment, the percentage of pupils in public sector schools, the percentage of the district that is urban, and a variable to indicate whether it is a city or county district. Interaction terms were also included for several of these school and district variables with some of the pupil-level variables.

Their value-added model was found to explain from 54 to 62 per cent of the variation in pupil test scores and from 59 to 80 per cent of the variation amongst schools. The pupil-level prior attainment in the third grade reading score was found to be significant at the 5 per cent level in positively influencing not only the pupil-level fourth-grade reading score, but also the fourth-grade maths score and the fourth-grade combined reading and maths score. The third-grade maths score was found to be only significant in positively influencing the fourth- grade maths score and the fourth-grade combined reading and maths score. Pupil-level African American ethnicity was significant at the 5 per cent level in negatively influencing the fourth-grade reading and combined score, though not the maths score. However, pupil-level African American ethnicity had a significant negative influence on all three scores when included in an interaction term with the percentage of the district that was urban. The age of the pupil was significant in negatively affecting all three scores. While per capita family income in the school district appeared insignificant, both the percentage of adults locally with 16 or more years of schooling, and the total size of the district enrolment, had significant positive influences at the 5 per cent level on all three scores, with students eligible for free or reduced price lunches having a similar negative influence.

The teacher test scores had a significant impact on the reading score, but not on the maths or combined score. The maths score was, however, significantly positively influenced by the proportion of teachers who had a Master's degree. Class size was specified as a series of dummy variables for different ranges of class size, several of which were found to be statistically significant in the overall value-added analysis, and with each estimated coefficient on the different class size variable taken to represent "the difference in student test scores relative to the base size of over twenty-nine students". They conclude that for fourth-grade maths "class sizes of under nineteen generate student test scores that exceed those for the base by 0.14 standard deviations. For reading, the estimated effects are somewhat smaller, about 0.05 standard deviations, and for the combined scores about 0.10 standard deviations. The reading gains were found to level off at a class size in the mid-twenties. But for mathematics and for the combined scores, gains in test scores occur throughout the relevant range of reductions in class size. More learning apparently occurs in smaller fourth grade classes than in larger classes, especially in math. Further investigation of the math gains indicates that the gains from smaller classes are greater for girls than for boys; a class size of less than nineteen

increases girls' math scores by 0.17 standard deviations relative to the base case, while a similar class size increases boys' scores by only 0.10 standard deviations" (*ibid*, p. 279-80).

d. School processes

In view of Raudenbush and Bryk (1989)'s emphasis on the desirability of including **school policy variables** as determinants of school effectiveness when data on these are available, it is of interest to note that several school effectiveness multilevel studies have included variables relating to educational practice within schools as explanatory variables. In a study of pupil achievement in science in Israeli elementary schools, Zuzovsky and Aitkin (1991) found that the degree of implementation of the new science curriculum at the school level, and related activities, such as teacher participation in in-service training for more than 10 hours a year, had a significant impact on pupil achievements in science. However, their study made use of pupil current reading scores, rather than prior attainment, as a proxy for pupil ability. In a value-added model which did take prior attainment explicitly into account, Plewis (1991) found that the extent of the mathematics curriculum which class teachers covered was linked, though not strongly, to the mathematics progress of infants in inner London schools. Gamoran (1991) tested two versions of the way in which educational practice may impact upon pupil value added in a multilevel model of the form:

$$y_{ij} = \alpha_j + \beta_j x_{1ij} + \gamma_j x_{2ij} + \delta x_{3ij} + \varepsilon_{ij} \qquad \text{for } i = 1, ..., n_j, \quad j = 1,...,m \tag{8.1}$$

where x_{1i} is a measure of pupil ability, as proxied by their prior attainment in an earlier relevant test, x_{2ij} is a measure of pupil effort and willingness to work, as reflected in pupil responses to an attitudinal questionaire, and x_{3ij} is a pupil-level socio-economic background variable. The **additive version** of (8.1) involves β_j and γ_j constant across schools, but school practices, as reflected in a school-level variable z_j , boosting the mean of the random intercept term α_j in (8.1), through the relationship:

$$\alpha_j = \alpha + \chi_o z_j + \varphi_{oj} \tag{8.2}$$

where φ_{oj} is the stochastic element of the school effect that is assumed to be normally distributed across schools, with a zero mean. In contrast, the **interactive version** of (8.1)

involves the intercept α_j constant across schools. Instead, it assumes that school practices provide individual pupils with the opportunity to raise the extent to which they can convert pupil ability and effort into pupil achievement, through the school-level variable z_j boosting the parameters β_j and γ_j on pupil ability and effort in (8.1), such that:

$$\beta_j = \beta + \chi_1 z_j + \varphi_{1j} \quad (8.3)$$

$$\gamma_j = \gamma + \chi_2 z_j + \varphi_{2j} \quad (8.4)$$

where φ_{1j} and φ_{2j} are stochastic and independently normally distributed with zero means.

When these versions were tested on two US data sets of eighth grade test results in mathematics and English respectively, Gamoran (1991) found that the results from both data sets were consistent with the additive approach, with the school variables of more content coverage in mathematics and a 'higher quality of instructional discourse' in English positively related to pupil progress in their respective subjects. However, the interactive version found support only in the data set for mathematics and only in boosting the coefficient on pupil prior attainment.

Based upon their survey of the literature on school and teacher effectiveness, including that based on value added studies, Sammons, Hillman and Mortimore (1995) identified a list key school and teaching variables which they consider to be the '**key characteristics of effective schools**'. These include (a) professional educational leadership within the school that is 'firm and purposeful' and encourages the participation of others in the management of the school participative; (b) shared vision and goals, unity of purpose, consistency of practice and collaboration and collegiality amongst staff within the school; (c) the school being an orderly and attractive working and learning environment; (d) concentration on teaching and learning through the maximisation of learning time and an emphasis and focus on academic achievement; (e) purposive teaching through structured lessons, adaptive practice, and well-prepared teaching with a clear purpose; (f) high expectations for all pupils that are clearly communicated to them and which provide them with intellectual challenge; (g) positive reinforcement of these expectations through discipline and feedback to pupils; (h) the monitoring and evaluation of pupil and school progress; (i) raising pupil self-esteem through

teacher attitudes and giving pupils responsibility and greater control over their work; (j) supportive home-school partnerships and parental involvement; and (k) school-based staff development and emphasis on the school as itself being a learning organisation. A similar range of factors contributing towards the process of school effectiveness is identified in the review of Reynolds and Teddlie (2000).

Hopkins and Reynolds (2001) have stressed the importance of the '**context specificity**' of the **appropriate school practices** that can boost a school's effectiveness and value added. Teddlie *et al* (2000) provide a review of the literature on such context-specific effective school practices. Borich (1996) highlights the different teacher behaviours which are likely to be effective for pupils in different socio-economic status (SES) areas. Differences between the school practices which were associated with effective US schools in middle socio-economic status areas from those in equally effective US schools located in low socio-economic status areas have also been identified by Teddlie and Stringfield (1993). These included the greater emphasis placed on external academic rewards in low SES schools, which tended to focus more on basic skills rather than an expanded curriculum, with headteachers who tended to be initiators rather than simply effective managers. Reynolds *et al* (2001) provide a review of the literature on school improvement and school practice that is particularly relevant to British schools in '**challenging circumstances**', with Stoll and Myers (1998) providing a discussion of the problems of boosting the effectiveness of schools that are initially in difficulty with their educational performance.

e. Educational policy initiatives

Many educational policy initiatives have as their motivation the raising of the effectiveness of schools in boosting pupil progress. One such initiative that has been subject to detailed evaluation using multilevel modelling is the former Technical and Vocational Education Initiative (TVEI) which sought to find ways of making education more vocationally relevant to 14- to 18-year olds across the ability range. In his study of the impact of TVEI in Scotland, Raffe (1991) used a TVEI dummy variable whose coefficient could vary across schools, alongside a pupil-level control variable to allow for the generally lower level of ability of students on TVEI who were on projects. He found a generally favourable effect of TVEI on self-reported pupil truancy, but not on educational achievement, participation in post-

compulsory education or employment, although the impact of TVEI varied across schools and individual TVEI projects. While pupil-level prior attainment test data were not available, Raffe used teachers' ratings of individual pupils' potential attainment assessed at the age of 14 as a proxy for pupil ability. When this variable was omitted from the analysis, there was not only a downward bias on the estimated TVEI effect but also a greater observed variability of the effect across schools and TVEI projects, which Raffe attributes to wide differences in the pupil intake across different TVEI projects.

f. Stability and time trends

An issue which arises in the context of the interpretation of the estimated school effect in value-added analysis is how far the school-level residual is genuinely a reflection of **differential school effectiveness** and how far it is instead the result of **random or other factors**, such as those related to pupil background, which have not adequately been taken into account (see Fitz-Gibbon, 1991). Thus, while both pupil background variables and the school effect may each account for only a relatively small part of the variance explained, once pupil prior attainment has been included as an explanatory variable, the omission of some relevant pupil background variables may lead to a proportionately large change in the estimated school effect.

If the estimated school effect in a given year is genuinely the result of other random factors, rather than of other variables that are more stable over time, one might expect a high degree of variation in the estimates of the individual school effects between years. In a value added study of 34 secondary schools in a single LEA, Gray, Goldstein and Jesson (1996) compared the levels and changes in the estimated school effectiveness for five separate successive cohorts of pupils who sat GCSEs over the period 1990-1994 in order to assess the extent to which individual schools improved their effectiveness over time. While there was a general upward trend in GCSE point scores, both nationally and across these 34 schools, over this period, there was also a significant variation across these 34 schools in the rate at which individual schools improved their average GCSE point scores. When a value-added multilevel analysis was carried out, there was found to be evidence of changes in individual school effectiveness over time. Although the correlation between the relative school effectiveness of individual schools in the years 1990 and 1991 was 0.88, the correlation between their scores for the years 1990

and 1993 dropped to only 0.52, so that the relative value-added scores of individual schools were not constant over this period.

Gray, Goldstein and Jesson (1996) also carried out a three-level analysis using year cohorts within each school as an intermediate level between pupils and schools, with both the year effect and pupil prior achievement effect allowed to vary randomly across schools . They found that 17 per cent of the between-schools variation in pupil value added could be attributed to the year cohort component. Thus, while there was a fair degree of stability in school value-added scores over the period, there was also a degree of change in them. The year effect for each school generated a coefficient that they interpreted as the **annual rate of improvement** for each school. When the initial level of effectiveness in 1990 and the yearly rate of improvement for each school were looked at together, there were found to be 6 schools out of the 34 with statistically significant positive value added, 5 schools with statistically significant negative value added, with the remaining two thirds of schools not departing significantly from their expected pupil outcomes, given their pupil prior attainments. At the same time there were found to be 10 schools that had statistically significant year effects, divided equally between those which were 'improving fairly rapidly' (i.e. about 1 in 7 of the schools) and those which were 'improving more slowly'. Both some more effective schools and some less effective schools were found to improve fairly rapidly over time, with some more effective schools and some less effective schools found to improve more slowly. The correlation coefficient between a school's initial value-added score and its annual rate of improvement was positive, but not large at 0.26. In general there was a tendency for schools which were initially found to be less effective to also be improving more slowly. Part, though not all, of the explanation for schools which were improving more rapidly was found to be a tendency by them to increase the number of GCSE examinations for which their pupils were entered.

Gray, Goldstein and Thomas (2001) carried out a similar value-added comparison over four successive years for pupil-level achievement at A/AS level across England as a whole using pupil prior attainments at GCSE, gender and institutional type as explanatory variables. They found that, while the **cross-years' correlations** in estimated individual school effects were of the order of 0.90 when pupil prior attainment was omitted as an explanatory variable, when it was included these correlations dropped to around 0.75 for adjacent years, to around 0.62 for cohorts that were two years apart and to only 0.55 for years that were three years apart.

As Sammons *et al* (1996) note, Raudenbush (1989b) and Nutall *et al* (1989) have stressed the desirability of adopting a longitudinal model for estimating school effects and their stability, when longitudinal data is available. If the school effect is decomposed into an average effect over a period of time and a time specific effect, Raudenbush (1989b) argued from his analysis of Scottish data that school effects on overall attainment are more stable than they first appear. However, Willms and Raudenbush (1989) found less stability prevailing in the school effect for specific subjects, a finding reflected elsewhere in the relatively few studies of this subject specific school effect stability (Sammons *et al*, 1996) .

Changes between years in the estimated school effect may result, however, not only from random factors around a stable average effect, but also from genuine **changes in the relative effectiveness** of different schools over time. Such changes may themselves result from factors such as a new headteacher or other staff changes, as well as changes in school practices, attitudes and priorities within the school. They may also result from a failure by the particular school to keep up with the improvements over time in examination results which are achieved nationally within the sample that is used to estimate their relative effectiveness.

The number of schools which exhibited patterns of **sustained improvements or sustained deteriorations** in their estimated value added over the whole of a 4-5 year period of time was found by Gray, Goldstein and Jesson (1996), Gray *et al* (1995), Thomas (2001), and Gray, Goldstein and Thomas (2001), to be relatively small. The lack of sustained improvements or deteriorations may itself result from a **negative feedback mechanism** whereby the management tends to relax if it initially appears to be performing well, but comes under greater pressure to improve if its performance shows initial deterioration. In addition, the lack of sustained improvements in school effectiveness may reflect pressure to make feasible improvements in school practices straight away to boost the initial school effectiveness after a change in the school management, rather than spread them out over a period of 4-5 years.

Pugh and Mangan (2003) argue that even where there appear to be consistent upward or downward trends over 4-5 years in a school's effectiveness, these may still result from random factors if the relevant model of its behaviour over time is not a deterministic trend-stationary process (with random variations around a deterministic linear time trend over time), but instead a **random walk with drift** (involving a difference-stationary stochastic process in which the

school effective at time t equals its value at time t-1 plus a constant (drift) term plus a white noise random error term).

Such a random walk process implies that past levels of the random disturbance terms exhibit a powerful influence over the expected current level of school effectiveness. This itself may imply that schools with an initially low level of school effectiveness because of their particular past history will tend to be difficult to 'turn around', with **cumulative self-reinforcing process** at work (see Stoll and Myers, 1998; Mayston, 2007a). While any subsequent sustained improvement in their estimated effectiveness may be due to stochastic factors, it may also be due to a superior 'drift' factor that results from improved management of the school. The ability to distinguish the two is likely to be aided by a close association between the value-added analysis and the identification of changes in the management and practices of the school which are likely to bring about improvements in its value added.

Significant differences in the rate of improvement in school GCSE scores over the period 1991-98 between schools in different circumstances were found in a study of over 300 secondary schools by Levacic and Woods (2002a). **School improvement rates** were found to be influenced less by social disadvantage *per se* than by **local school hierarchies.** High rates of improvement were found to be associated with **low concentrations of social disadvantage** for a school compared to other local schools. Although there was greater scope for improvement due to starting from a relatively low base of results, schools that had high rates of social disadvantage relative to other schools suffered **low rates of improvement** in their GCSE scores. There was also a tendency for these schools to find that their **relative social disadvantage increased over time in competition with more successful local schools**, who also tended to have the advantage of benefiting financially from a greater rate of improvement and improved pupil demand. In a follow-up case study of three of the schools who had substantially lower GCSE scores than the national average, Levacic and Woods (2002b) examined the greater internal and external barriers to improvement which socially disadvantaged schools face. These include the **cumulative impact** of within-school effects of a more difficult and disrupted learning environment. They also included management processes that proved less able to cope with these problems, and the cumulative contextual and school performance effects that reduce the **local competitive position** of the schools and its perceived attractiveness to pupils and their parents. While their analysis was carried out before more recent targeted initiatives, such as Education Action Zones and Excellence in Cities, were

underway and they did not carry out a multilevel value-added analysis to adjust the changes in GCSE scores for differences in pupil prior attainments, the findings of Levacic and Woods (2002a,b) reinforce the need to pay attention to the particular problems which face socially disadvantaged schools in seeking to improve their GCSE performance.

g. Inspections

Based on a logistic regression for a stratified sample of schools, Cullingford and Daniels (1999) claimed to find a significant negative impact of OFSTED inspections on the subsequent proportions of 5 or more A* - C grade GCSEs obtained relative to non-inspected schools. Using a multilevel analysis of GCSE results, Shaw *et al* (2003) find that 'inspection had a consistent, negative effect on achievement, depressing it by about one half of a per cent' in the case of county mixed comprehensive schools. On contrast, inspection was found to increase achievement by about one per cent in grant-maintained mixed comprehensives, and had a mixed effect for other types of school. However, a value-added approach was not adopted in either study, with no adjustment made for pupil prior attainment or other socio-economic context variables in estimating an inspection effect.

h. Contextual Value Added

In addition to pupil prior attainment, several groups of variables have been used in the Contextual Value Added models produced by the DfES (2005) to take into account the additional pupil- and school-level factors which may influence the rate of pupil progress. For pupil progress between KS2 and KS4, between KS3 and KS4 and between KS2 and KS3, the additional pupil-level variables include the pupil's gender, their eligibility for Free School Meals, their Special Educational Needs status, their ethnicity, their age, whether they have been in care whilst at the current schools, whether they have moved between schools at non-standard transfer times, and a measure of deprivation given by the Income Deprivation Affecting Children Index (IDAC) score based upon the pupil's postcode. The school-level variables include the average prior attainment score and the standard deviation of this score for the relevant cohort of pupils in the school. However, in the estimation of the extent to which these additional pupil- and school-level variable contribute towards the prediction of pupil performance, DfES (2005) notes that "even when we include contextual factors we find prior attainment is by far the strongest predictor of outcomes".

In contrast to the earlier reliance of the published value-added scores on the levels which pupils achieved in the relevant Key Stage, the point scores at KS2 and KS3 used to assess pupil Contextual Value Added use 'fine grades' based upon the underlying marks which pupil achieve at these Key Stages. As noted in Section 6 above, the earlier reliance upon levels produced **discrete jumps** in the point scores, for instance from 27 points for level 4 to 33 points for level 5, in response to possibly smaller proportionate changes in the underlying marks. Pupil who achieve the minimum mark for level 4 will now instead be assigned 24.0 points, while a pupil attaining marks at the mid point between the level 4 and 5 boundaries will be assigned 27.0 points, and a pupil who achieves marks just below the level 5 threshold will be awarded nearly 30.0 points. The 'fine grade' average point score in English, Mathematics and Science is used to reflect pupil prior attainment at KS2 or KS3, together with the square of the average point score, to take account of possible non-linearities in the influence of such prior attainment on pupil performance. The differences between the pupil's prior attainment in English and in Mathematics from their overall average point score, using such 'fine grades', are also included as pupil prior attainment variables.

9. THE FUNCTIONAL FORM

A further consideration which may significantly affect the value-added assessments of individual schools, and particularly those who are **outliers in the overall distribution**, is the functional form adopted for the inclusion of the independent and dependent variables in the regression equation for determining each pupil's predicted outcome scores. Trower and Vincent (1995), for example, found a cubic functional form to provide the best fit in their analysis of value added in secondary schools. However, except when used as local approximations, quadratic and cubic functional forms can have the counter-intuitive implications of expected pupil achievements over some range being a decreasing function of pupil prior attainment or of the other explanatory variables to which they are applied.

a. An economic formulation

One possible functional form for the production function for educational value added that can bring together resource and other inputs alongside pupil prior attainment and other pupil characteristics, and which ties in with economic analysis, is that suggested in Mayston (2007b). This involves a pupil-level Cobb-Douglas educational production function of the form:

$$y_{i\ell hj} = A_{\ell hj} . F(m_{\ell hj}) . \prod_{k \varepsilon S} x_{ki\ell hj}^{\alpha_{kj}} . \prod_{s \varepsilon G} v_{si\ell hj}^{\beta_{sj}} \qquad (9.1)$$

where $y_{i\ell hj}$ denotes the level of educational outcome in subject h for pupil i in class ℓ in school j at a given stage in the educational process, such as GCSE. $m_{\ell hj}$ is the class size for class ℓ in school j for subject h. The variable $x_{ki\ell hj}$ denotes the level per pupil of resource k for pupil i in class ℓ in school h for subject j. S is the set of resource inputs deployed. The variable $v_{si\ell hj}$ denotes the level of pupil characteristic s for pupil i in class ℓ in school h for subject j. G is the set of these pupil characteristics, such as ethnicity and socio-economic background variables. Also included in G is the level of pupil prior attainment, which we will designate as $v_{oi\ell hj}$, at the appropriate earlier in the educational process, such as at KS3 or KS2. The parameters α_{kj} for all k in S, and β_{sj} for all s in G are assumed constant for a given subject j. Each subject is differentiated by the level or Key Stage at which it takes place, so that GCSE mathematics clearly refers to a different educational output to KS3 mathematics.

The resource variables, $x_{ki\ell hj}$ for all k in S, the set of all pupil-level resource variables, are measured in real physical terms, such as the quantity of teaching time received, per pupil i in class ℓ in school j for subject h. The class size variable, $m_{\ell hj}$, is included to permit considerations of the extent of returns to scale in class size, as reflected in the function F. These economies of scale may vary as the class size expands, reflecting the extent to which educational output will be improved by teaching pupils together in a class of more than one pupil, for a given level of teaching time and other resources *per pupil*. Thus a class size of 20 in which 20 pupils spends an hour together with one teacher may be more educationally productive than twenty class sizes of one, in which each pupil receives only three minutes of teaching time. However, as the class size expands diseconomies of scale are likely to set in with a wider degree of heterogeneity of pupils in the class.

The term $A_{\ell hj}$ in equation (9.1) is a measure of the overall educational effectiveness of class ℓ in school j in subject h. Its overall educational effectiveness may result in part from class-level resources, such as the qualifications of the class teacher(s), and in part from class-level process variables, such as the number of disruptive pupils in the class, as well as from the overall educational effectiveness of school j in subject h. We can denote by $y_{k\ell hj}$ the quantity of class-level resource k that is deployed in class ℓ in school j for subject h, for all k in T, where T is the set of all such class-level resource variables (excluding class size, which is already included in (9.1)). We can denote by $u_{s\ell hj}$ the magnitude of the class-level process variable s, for s ε Z, where Z is the set of all such class-level process variables. We may then postulate a relationship of the form:

$$A_{\ell hj} = B_{hj} . \prod_{k \varepsilon T} y_{k\ell hj}^{a_{kj}} . \prod_{s \varepsilon Z} u_{s\ell hj}^{b_{sj}} \qquad (9.2)$$

where the parameters a_{kj} and b_{sj} are assumed to be constant, and where B_{hj} is a measure of the overall school-level effectiveness of school h in subject j.

The school-level effectiveness term B_{hj} may itself reflect in part school-level resource variables, such as the qualifications of the headteacher, and school-level process variables, such as whether or not the school has been re-organised in the last five years, as well as national subject-specific factors, such as the content of the National Curriculum in that subject. We will denote by z_{khj} the magnitude of the kth school-level resource variable in school j for

subject h, for all k in N, the set of all such school-level resource variables. Similarly, we can denote by w_{shj} the magnitude of the sth school-level process variable in school j for subject h, for all s in V, the set of all such school-level resource variables. We can then postulate that:

$$B_{hj} = D_h . \prod_{k \varepsilon N} z_{khj}^{c_{kj}} . \prod_{s \varepsilon V} w_{shj}^{d_{sj}} \qquad (9.3)$$

where the parameters c_{kj} and d_{sj} are assumed to be constant, and where D_h is a parameter that reflects factors affecting the overall national educational effectiveness in subject h beyond the resourcing and process variables that have already been included in the analysis.

As emphasised in Mayston (2003a), the inclusion of resource variables into the analysis enables issues of **value for money** and **cost-effectiveness** to be considered together with issues of educational effectiveness. However, as Mayston and Jesson (1999) stressed, their inclusion and effective deployment within the empirical analysis is dependent upon adequate data being available on resource use within schools in a comprehensive and comparable form. The Consistent Financial Reporting (CFR) framework that has since been developed by the DfES (2006) has made considerable progress in developing a comparable database for different school-level resource expenditure categories. More detailed comparative data on how resources, such as teachers' time, are deployed across different pupils and subjects are not currently available, despite the potential scope for electronic timetabling systems to generate such data.

Equations (9.1) – (9.3) have the Cobb-Douglas property of being linear in the logarithms of the respective resource and non-resource variables. Equation (9.1), for instance, can be transformed into the form:

$$\ln y_{i\ell hj} = \ln A_{\ell hj} + \ln F(m_{\ell hj}) + \sum_{k \varepsilon S} \alpha_{kj} \ln x_{ki\ell hj} + \sum_{s \varepsilon G} \beta_{sj} \ln v_{si\ell hj} \qquad (9.4)$$

where ln denotes a natural logarithm, and similarly for equations (9.2) and (9.3).

The functional relationships (9.1) – (9.4) describe the efficient points on the *educational production possibility frontier*, of the maximum possible levels of pupil attainment, given the pupil and school resource and other inputs. Associated with (9.1) and (9.4) are the regression-based relationships:

$$\ln y_{i\ell hj} = \ln A_{\ell hj} + a_j.m_{\ell hj} + m^2_{\ell hj} + \sum_{k\varepsilon S} \alpha_{kj} \ln x_{ki\ell hj} + \sum_{s\varepsilon G} \beta_{sj} \ln v_{si\ell hj} + e_{i\ell hj} \quad (9.5)$$

where

$$\ln A_{\ell hj} = \ln B_{hj} + \sum_{k\varepsilon T} a_{kj} \ln y_{k\ell hj} + \sum_{s\varepsilon Z} b_{sj} \ln u_{s\ell hj} + \theta_{\ell hj} \quad (9.6)$$

and

$$\ln B_{hj} = \ln D_h + \sum_{k\varepsilon N} c_{kj} \ln z_{khj} + \sum_{s\varepsilon V} d_{sj} \ln w_{shj} + \mu_{hj} \quad (9.7)$$

for the choice of $F(m_{\ell hj}) = \exp(a_j.m_{\ell hj}^2 + b_j.m_{\ell hj})$.

The disturbance terms in (9.5) – (9.7) are made up of three parts for each subject j, with each disturbance term assumed to have an expected value of zero. The first, μ_{hj}, at the school level in equation (9.7), reflects school j's relative effectiveness in subject h at the school level. It will be positive for those schools which have a higher level of overall school effectiveness in (9.7) than that predicted by the regression analysis, and negative for those schools which have a lower level of overall school effectiveness than that predicted by the regression analysis. The second disturbance term, $\theta_{\ell hj}$, in equation (9.6), reflects relative effectiveness at the class level. It will be positive for those classes which have a higher level of overall class effectiveness in (9.6) than that predicted by the regression analysis, and negative for those classes which have a lower level of overall class effectiveness than that predicted by the regression analysis. The third disturbance term, $e_{i\ell hj}$, at pupil level in equation (9.5) reflects variations across individual pupils which are not wholly captured by the other terms in (9.6) and (9.7), and which are not accounted for by pupil characteristics and by pupil-level resource variations. If the dataset were rich enough, (9.5) – (9.7) could readily be extended to include year group and also LEA-level effectiveness terms.

Insertion of equation (9.7) into equation (9.6), and equation (9.6) then into equation (9.5), yields a multilevel model of the determination of pupil-level educational performance, based upon an error component formulation in which the overall disturbance term is made up of the above three disturbance term components. As noted in Section 11 below, issues arise in this context as to the extent to which some process variables, such as those related to pupil attitudes and school expectations, are endogenous to the overall level of school resourcing and educational performance.

Associated with the educational production function (9.1), we may write a **pupil-level cost function**:

$$C_{i\ell hj}(p,\, y_{i\ell hj}) \;\equiv\; \min_{x_{ki\ell hj}} \sum_{k\varepsilon S} p_{kj}.x_{ki\ell hj} \quad s.t. \quad A_{\ell hj}.F(m_{\ell hj}).\prod_{k\varepsilon S} x_{ki\ell hj}^{\alpha_{kj}}.\prod_{s\varepsilon G} v_{si\ell hj}^{\beta_{sj}} \;\geq\; y_{i\ell hj} \qquad (9.8)$$

where p_{kj} is the price of input k that school j faces, with p the associated price vector of all such prices. This analysis permits variations in the cost of attracting teachers of a given level of quality across different schools and localities.

Under the assumption that the prices, p_{kh} , which each school h faces do not vary with the level of its resource use, we may derive the pupil-level cost function associated with the educational production function (9.1) to be:

$$C_{i\ell hj} = (\alpha_j.y_{i\ell hj} / A_{\ell hj})^{1/\alpha_j}.\prod_{k\varepsilon S}(p_{kj}/\alpha_{kj})^{\alpha_{kj}/\alpha_j}.\prod_{s\varepsilon G} v_{si\ell hj}^{-\beta_{hj}/\alpha_j}.[F(m_{\ell hj})]^{-1/\alpha_j} \; where \; \alpha_j \equiv \sum_{k\varepsilon S}\alpha_{kj} \qquad (9.9)$$

involving the parameters $A_{\ell hj}$, α_{kj}, β_{hj} and of F which may be estimated from the regression relationship (9.5) – (9.7). In addition, (9.9) involves the pupil characteristics $v_{si\ell hj}$, including the pupil prior attainment level $v_{oi\ell hj}$, together with the local prices p_{kj} . It is important to note that the cost function (9.9) will not in general imply constant unit costs per pupil or per unit of the educational output. The approach of estimating the parameters of an educational production function in (9.1) – (9.3) may well then differ from that involved in a costing model based upon the use of average costs per pupil that are implicitly assumed to be constant over the range of variation in the sample. From (9.9) we can also derive the inverse function:

$$y_{i\ell hj} = R_{i\ell hj}^{\alpha_j}.(A_{\ell hj}/\alpha_j)^{1/\alpha_j}.\prod_{k\varepsilon S}(\alpha_{kj}/p_{kj})^{\alpha_{kj}}.\prod_{s\varepsilon G} v_{si\ell hj}^{\beta_{sj}}.[F(m_{\ell hj})] \qquad (9.10)$$

that specifies the attainable level of the educational output $y_{i\ell hj}$ that can be obtained from a given level of pupil-level resource expenditure, $R_{i\ell hj} = C_{i\ell hj}$, given the other parameters involved. While costs and resource expenditure are measured here in money terms, equation (9.10) involves a form of price deflator to adjust the overall monetary expenditure variable $R_{i\ell hj}$ to real terms by taking into account the level of local inputs prices p_{kj} .

b. Other functional forms

A further alternative functional form for the educational function, to the standard linear equations of multi-level modelling, which has been estimated empirically is that proposed by Montmarquette and Mahseredjian (1985, 1989). This involves estimating a logit function of the form:

$$y_i = 1/[1 + \exp(-\sum_k \beta_k x_{ki})] \qquad (9.11)$$

where each individual pupil examination score, y_i , is now restricted to range between 0 and 100 per cent. This is equivalent to replacing each output variable, y_i , by the following **logistic transformation** in the relationship:

$$q_i \equiv \ln[y_i /(1 - y_i)] = \sum_k \beta_k x_{ki} \qquad \text{for } 1 > y_i > 0 \qquad (9.12)$$

In contrast to the standard linear form of the educational production function, (9.11) implies that y_i does lie within the bounds of the feasible range for the examination performance under consideration. If a point score not involving 0 to 100 per cent is used, such as for GCSE grade point scores that exceed 1.0, it can be readily transformed to one that does range from 0 to 100 per cent by dividing by the maximum possible point score.

Moreover, rather than following the linear assumption that the marginal influence of each of the explanatory variables is constant across the whole range of pupil achievement, (9.12) involves a non-linear S-shaped curve where the marginal impact of each explanatory variable x_{ki} is greater around the **inflection point** at a 50 per cent points score y_i than at very high or very low values of y_i . By adopting an inverse power transformation of (9.11), Montmarquette and Mahseredjian (1985) generalise this S-curve to have an inflection point that is determined by the data. In their empirical analysis of Montreal schools, Montmarquette and Mahseredjian (1985) found, using Davidson and MacKinnon's (1981) model specification tests, that this transformation provided an improved model specification to the linear model. Using non-linear estimation techniques, they estimated their inflection point at 68 per cent, and found using this

transformation that variables such as class size, mother's education and pupil IQ had a larger estimated influence using the transformed value of pupil achievement than they did in a linear specification of the model.

Another possible functional form for the educational production function and which "has proved the most popular form in recent applied production economics" (Chambers, 1988, p. 168) is the **translog** function. This involves a quadratic functional form in the logarithms of the m underlying input variables, of the form:

$$\ln y_i = \beta_o + \sum_{k=1}^{m} \beta_{ki} \ln x_{ki} + 0.5 \sum_{k=1}^{m} \sum_{h=1}^{m} \gamma_{kh} \ln x_{ki} \ln x_{hi} \qquad (9.13)$$

where y_i is the examination score of pupil i and x_{ki} is the value of the kth input variable, such as prior attainment, for pupil i. The translog function (9.13) involves a **flexible functional form** that provides a second-order local numerical approximation to a general production function. However (9.13) also involves estimating a potentially large number of additional parameters, namely the m(m+1)/2 distinct cross-parameters γ_{kh}. The Cobb-Douglas production function discussed above is a special linear form of (9.13) in which each γ_{kh} is assumed to be zero.

Although the cross-terms could be incorporated into a multilevel model using **interaction effects**, their use in a value added context has not been widespread in published papers. Aitkin and Zuzovsky (1992) argue that interactions may be fitted more powerfully within a multilevel model by deleting the main effects from the model in order to reduce a major source of multicollinearity that tends to increase the standard error of the interaction terms.

Other possible functional forms for the educational production function include those which are linear in transformed output and input variables, using Box and Cox (1964) transformations of the form:

$$y(\lambda_o) = (y^{\lambda_o} - 1)/\lambda_o \text{ for } \lambda_o \neq 0, \; y(\lambda_o) = \ln y \text{ for } \lambda_o = 0 \qquad (9.14)$$

$$x_k(\lambda_k) = (x^{\lambda_k} - 1)/\lambda_k \; for \lambda \neq 0, x_k(\lambda_k) = \ln x_k \; for \lambda_k = 0 \qquad (9.15)$$

Cases of interest include the **semi-log** case in which $\lambda_o = 1$ and $\lambda_k = 0$ for all k = 1, .., m. Here each input variable, such as pupil prior attainment, has a diminishing marginal impact on a pupil's educational achievement, but which may also result in a potentially negative value to examination performance. A further case of interest is that of $\lambda_o = 0$ and $\lambda_k = -1$ for $x_k > 0$, corresponding to the **logarithmic-reciprocal model** (see Johnston, 1987, p. 71) in which the natural logarithm of examination performance is a decreasing linear function of the reciprocal of the input variable x_k. This functional form takes on a similar shape to the logistic curve, with examination performance flattening out asymptotically towards its maximum value as the input variable, such as prior attainment, increases to high values, but rises steadily as the input variable increases over an intermediate range, and examination performance falls more slowly towards zero as the input variable declines towards zero. The logarithmic-reciprocal curve can handle cases where y is 100 per cent, though not zero. The logistic curve requires use of maximum likelihood estimation methods if zero and 100 per cent values to y are possible (see Gujarati, 1995, p. 556).

c. Stochastic frontier analysis

In conventional value-added analysis, as in equation (3.4) above, the stochastic disturbance terms are typically taken to be jointly normally distributed (see Goldstein, 1995, p. 22), with the school effect often interpreted as a measure of the school's educational effectiveness. The logarithmic formulation in (9.6) and (9.7) above fortunately avoids the undesirable implication of possible negative values to examination results that such normality would imply. An approach which would not necessarily attribute the stochastic disturbance term at the school level to simply variations in the effectiveness of individual schools is provided by stochastic frontier analysis. As proposed by Aigner, Lovell and Schmidt (1977), stochastic frontier analysis involves decomposing the stochastic disturbance term into two parts. The first involves a random disturbance term v_j that is assumed to be identically, normally and independently distributed across all producers j, and reflects inter-firm **heterogeneity** due to underlying random elements that are not adequately taken into account by the explanatory variables of the measured inputs but which affect the position of the feasible production possibility frontier for the producer. The second involves a non-negative disturbance term u_j that is a measure of the extent of the inefficiency of producer j, as reflected in the degree of

departure of their actual output from their maximum feasible output for their given inputs that is indicated by the production possibility frontier for the producer. The u_j are assumed to be identically independently distributed from a truncated normal distribution over such non-negative values, with the u_j and v_j terms independently distributed.

Within the multilevel value-added formulation given by (3.4) above, this would imply:

$$q_{ij} = \alpha + \beta x_{1ij} + \gamma x_{2ij} + \delta s_j + \theta_j + \varepsilon_{ij} \quad \text{where } \theta_j = v_j - u_j \text{ for } i = 1, ..., n_j ; j = 1,...,m \qquad (9.16)$$

where v_j and u_j have the above properties. Only the $-u_j$ term would then be taken to reflect school *j*'s educational effectiveness, with $-u_j$ being equal to zero for fully effective schools and becoming more negative as school *j*'s educational effectiveness declines. If $-u_j$ corresponds to the natural logarithm of an underlying multiplicative effectiveness term ς_j, as in a Cobb-Douglas formulation of the underlying educational stochastic production function, the school effectiveness indicator ς_j, will then lie between 1.0 for fully effective schools and zero for completely ineffective schools. If the term *value added* is interpreted as a measure of the educational effectiveness of the school relative to what the school could be expected to produce, given its underlying production possibilities, then it is $-u_j$ and the associated indicator ς_j which are relevant here. However, this requires the $-u_j$ term to be isolated from the heterogeneity term v_j, which itself may reflect additional factors, such as differences in the underlying quality of the resource inputs into the production process which the school has available to it.

Jondrow et al (1982) provide a method for separating out the two components v_j and u_j for each producer in a standard single-level analysis. Battese and Coelli (1995) extend the basic stochastic frontier model by assuming that the u_j efficiency terms are not identically distributed across producers, but have a mean which varies linearly within producer characteristics. Further extensions of the basic stochastic frontier model are discussed in Kumbhakar and Lovell (2000).

Greene (2004) demonstrates the importance of distinguishing the effects of heterogeneity between different producers from the assessment of producer efficiency in the context of

international comparisons of national health care systems. However, Greene (2005) shows that different approaches to incorporating such heterogeneity within stochastic frontier models can produce "very different results" (ibid, p. 298). The application of stochastic frontier analysis to multilevel models, and a comparison of different approaches to incorporating producer-level heterogeneity, in the context of assessing the effectiveness of different schools remains an area for further research.

d. Effectiveness evaluation and multiple outputs

A further important consideration in modelling the effectiveness of different schools, and the impact which educational programmes may have on this effectiveness, is the **multi-dimensional** nature of their outputs. Secondary schools in particular typically contribute towards educational attainment at more than one stage of the educational process, such as Key Stages 3, 4 and 5, and in several different subjects. One approach to assessing their effectiveness is to make separate assessments at each level and for each main subject category, such as English, Mathematics and Science. This approach is pursued in the formulation of the educational production function (9.1) above, and in the multilevel estimations of the DfES (2005) Contextual Value Added models discussed in Section 8 above.

However, an alternative approach would be to recognise more explicitly the **trade-offs** which may exist between what is attainable for the school in each of these different directions, particularly once resources are incorporated into the educational production function and the budgetary constraints which schools face are recognised. An analytical technique which can produce a summary estimate of the effectiveness of each individual school within this multiple output context is that of **Data Envelopment Analysis** (DEA) (see e.g. Mayston and Jesson, 1988; and Cooper, Seiford and Tome, 2000). The effectiveness measure which DEA produces is of its **technical efficiency**, which in the output-oriented case is inversely related to the proportion by which each of its outputs could be expanded (holding constant its existing output-mix), whilst still remaining in the production possibility set that DEA estimates based upon a convex hull of the input-output vectors of schools in the sample.

Data envelopment analysis is a non-parametric technique that avoids the need to specify a particular functional form for the underlying educational production function, although it still involves other important restrictive assumptions, such as convexity and homotheticity (see

Mayston, 2003). DEA also does not involve any stochastic structure and associated statistical estimation, making its efficiency estimates potentially very sensitive to data inaccuracies and measurement errors. A bootstrap procedure to estimate the sensitivity of DEA's technical efficiency estimates to stochastic variation of the efficiency frontier is developed by Simar and Wilson (2000). Sickle (2005) finds in the context of panel data analysis that DEA performs well compared to stochastic frontier and other estimators in a Monte Carlo simulation study of the impact on the efficiency estimates of a range of misspecifications of the underlying inter-temporal model.

As noted above, DEA's central concept of technical efficiency involves holding constant the producer's existing output mix. A similar concept of output-orientated 'radial' technical efficiency is indeed proposed in Fernandez et al (2005) for multiple-output production functions that are estimated using other techniques than DEA. In the case of a school, this would mean accepting the school's existing balance of achievements at different stages of the educational process, and existing balance of achievements between different subjects, such as English, Mathematics and Science. However, a key part of the effectiveness evaluation of a school's performance is likely to be questioning this balance of achievements and examining whether the school lags behind other schools in similar circumstances in each relevant direction of educational attainment. As we have noted in Section 7 above, individual schools may differ significantly in their relative effectiveness across different subjects and across different pupil groups. While Farrell (1957) originally considered the concept of **price efficiency**, in addition to his seminal contribution to the measurement of technical efficiency, the measurement of price efficiency would require here knowledge of the prices to be placed upon the different dimensions of educational attainment. In the absence of any such explicit prices, there is a strong case for examining the school's attainment in each relevant direction, as in the above CVA analysis, combined with an awareness that there may be trade-offs between what a school can achieve in each relevant direction.

10. DIFFERENTIAL SCHOOL EFFECTIVENESS

An important further extension of the basic value-added model of school effectiveness is of the following form:

$$y_{ij} = \alpha + (\beta + \xi_{1j})x_{1ij} + (\gamma + \xi_{2j})x_{2ij} + \delta s_j + \theta_j + \varepsilon_{ij} \qquad \text{for } i = 1, ..., n_j \text{ ; } j = 1,...,m \qquad (10.1)$$

where ξ_{1j} and ξ_{2j} are assumed to be independent **random school-level coefficients** that may still nevertheless be correlated with the school intercept term θ_j. The inclusion of such random coefficients permits the responsiveness of pupil achievement levels, y_{ij} , to different pupil prior attainment levels, x_{1ij} , and to other pupil-level variables, such as x_{2ij} , to vary across schools. Different schools may then exhibit different degrees of effectiveness in a way which depends upon the level of the pupil characteristics involved. One school may, for example, secure very high achievement levels for pupils with strong prior attainments, but very low results for pupils with weak prior attainments, whereas another school may secure less divergent results for its pupils. Extending the value-added model in this way enables recognition to be given to the fact that a school may be more educationally effective in adding value to some groups of pupils, but less effective for others. Issues of **equality of treatment and educational effectiveness across these different pupil groups** can also therefore be examined.

Using a form of (10.1) in which the pupil VRQ prior attainment score is the only included pupil-level explanatory, Aitkin and Longford (1986) found no significant differences between the schools in their sample in the slope coefficients on this pupil-level variable, once one exception school with a much higher slope coefficient was removed from the sample. Nuttall *et al* (1989) found significantly greater variability in the effectiveness of different schools amongst pupils with high levels of prior attainment than for pupils with low prior attainment levels. Similarly, Smith and Tomlinson (1989, p. 282) concluded that: "There are important differences between schools in the balance of their success as between below-average and above-average pupils. Nevertheless, the same schools achieve good and bad results both with below-average and with above-average pupils. There is more difference between the results achieved by different schools in the case of above-average than in the case of below-average pupils. In other words, both a below-average and an above-average child benefits from going to a good school: but the above-average child benefits more. This may be largely because the

exam system is such that the below-average child has little prospect of getting results however well he or she is taught".

Smith and Tomlinson (1989, p. 272) also concluded that the best fitting model was one where examination results in English were modelled separately and the coefficients on pupil prior attainment in the second year reading test and on the country of origin variables included a random school component. In this model, "the school level accounts for about 10 per cent of the variance where the second-year reading score was high or low and for about 2 per cent where it was average... Among pupils belonging to both broadly defined ethnic minority groups, the proportion of variance at the school level is considerably higher. This means that there are sharper differences between schools in rate of progress among ethnic minorities than among the white majority".

Trower and Vincent (1995) found in a **differential slope** version of the multilevel value-added model for 39 secondary schools' GCSE performance that the correlation between school residuals for pupils in lower quartile (labelled Q1) of KS3 scores and those in the upper quartile (labelled Q3) was 0.823, while the correlation between the school residuals for pupil in the lower quartile and those with the mean value of KS3 scores was 0.970. They drew the strong conclusion that: "It is clear that a school's residual at Q1 can be radically different from its residual at the mean or at Q3 and that the residuals at more than one point are need to give a complete picture", and that: "There was strong evidence that some schools have varying degrees of effectiveness with pupils of differing starting attainments; a school may be better than others with the more able pupils but do less well with its less able pupils. A single indicator could not, therefore, tell the whole story since there were schools which would be judged 'good' for pupils with one KS3 score but 'not so good' for pupils with another".

However, while differences in ethnic background may differentially affect progress in English, Gray, Jesson and Sime (1995) found 'little substantive evidence' for the existence of differential slopes in any of the datasets which they analysed in explaining overall pupil attainment in GCSE examinations. This is broadly in line with the findings of Jesson and Gray (1991) that "pupils of different prior attainment levels did slightly better in some schools than others". This contrasts with the conclusions of Nuttall *et al* (1989) that their research based on ILEA data "has found that school effectiveness varies in terms of the relative performance of different subgroups. To attempt to summarise school differences, even after adjusting for

intake, sex and t he ethnic background of the student and the fixed characteristics of the school in a single quantity is misleading". Jesson and Gray (1991) argue that this conclusion was partly due to the cruder data on pupil prior attainment that Nuttall *et al* (1989) employed, which were grouped into only three broad bands and tended to under-reflect the wide spectrum of pupil intakes which ILEA schools admitted. When a more finely differentiated prior attainment measure was used in a value-added analysis of the ILEA results, Jesson and Gray (1991) found that "the estimates for the slopes showed little evidence that they varied significantly".

Sammons *et al* (1993) examined the extent of differential school effectiveness for mathematics and reading in influencing pupil progress between entry in year 3 and at year 5 in a multilevel analysis of a longitudinal database for junior schools in the Inner London Education Authority. They found evidence of increases over time in the variance of pupils' achievements attributable to differences between schools, especially in mathematics, and marked differences between individual schools in their value added. In addition, they found "some evidence of differential effectiveness (differential slopes) of individual schools for pupils with different prior attainment levels. For reading, school differences were found to be greatest for pupils with low initial attainment. For mathematics, school differences were greater both for those with low and those with high initial attainment, and lower for those with average initial attainment". However, as Sammons (1999) confirms, they found no evidence of differential school effectiveness for different pupil groups differentiated by gender, ethnicity, FSM status or social class.

Thomas and Mortimore (1996) adopted a different approach to assessing the differential effectiveness of individual schools across different pupil groups. Instead of deploying the differential slopes formulation (10.1), they divided pupils into three groups according to whether they were 'high, low or average attainers on entry to school'. They concluded from their separate multilevel analyses of the school effects for these different groups of pupils that "schools are significantly differentially effective" across these three bands of prior attainment. Whilst the correlation between the school value-added scores for the highest and the average prior attainment bands was 0.82, it was found to be only 0.44 between the school value-added scores for the highest and the lowest prior attainment bands.

Thomas *et al* (1997) found differential school effects, that were significantly different at the 5 per cent level, in total at GCSE and in several individual subject scores, over different levels of prior attainment, across pupil gender groups, across Caribbean and other ethnic groups and according to whether pupils were entitled to FSM. They also found significant differences across schools in their trends in performance over the three years studied (1990-92) for English, English literature and mathematics, but not for the total GCSE score, French, history or science. In examining the consistency of the effectiveness of individual schools and subject departments, they conclude that "the results suggest that all pupils in effective schools and departments ... are likely to perform relatively well at GCSE but that particular groups of pupils (such as non-FSM pupils) are likely to perform especially well. In contrast, it appears that all pupils in less effective schools and departments ... are likely to perform relatively poorly at GCSE but that particular groups (such as 'other' ethnic groups) are likely to perform not quite so poorly".

11. MEASUREMENT ERRORS AND ENDOGENEITY ISSUES

A potential difficulty with the application of both OLS and multilevel regression approaches to the assessment of value added is the possible breach of one of the standard requirements for **unbiased estimates** to result from the application of these estimation techniques, namely that the explanatory variables that enter into the educational production function are **uncorrelated** with the residuals in the regression analysis for the school's educational performance. In this section, we examine two main ways in which such a correlation may occur. The first relates to the existence of errors in the measurement of the explanatory variables that enter into the educational production function. The second relates to the existence of other important inter-relationships between these explanatory variables and educational performance in addition to that described by the concept of the educational production function.

a. Measurement error

One interpretation of the concept of an educational production function might be as a relationship between individual pupil educational attainment at a given stage g of the education process and individual pupil ability, with school effectiveness in educating a given pupil in the school reflecting the extent to which the individual pupil does achieve the maximum level of educational attainment at the given stage g that their individual pupil ability implies that they are capable of achieving. Such an educational production function might be written in the form:

$$q_{ijg} = \alpha + \beta a_{ij} + \eta_{ijg} \quad \text{where} \quad \beta > 0 \tag{11.1}$$

where a_{ij} is a true measure of the ability of pupil i in school j, and η_{ijg} is an index of the effectiveness of school j in educating pupil i at stage g of the educational process, where both would be in logarithmic form in the case of a Cobb-Douglas production function. For conventional regression-based estimation, we may normalise the school effectiveness index by setting $E(\eta_{ijg}) = 0$.

In this context, pupil prior attainment at an earlier stage g-r of the educational process might well be an imperfect measure of the pupil's underlying ability. Using pupil prior attainment as

the explanatory variable in a regression aimed at estimating the relationship between pupil ability and pupil attainment at stage g will then involve an error in the measurement of the true underlying explanatory variable in the educational production function. We may then write:

$$q_{ijg-r} = a_{ij} + e_{ijg-r} \tag{11.2}$$

where e_{ijg-r} is an error term that is assumed to be uncorrelated with η_{ijg} and a_{ij}, and have a variance σ_e^2, and the individual ability measure a_{ij} is normalised such that $E(a_{ij}) = E(q_{ijg-r}) \equiv \mu_{g-r}$, and hence $E(e_{ijg-r}) = 0$. (11.1) - (11.2) imply that:

$$q_{ijg} = \alpha + \beta q_{ijg-r} + \varepsilon'_{ijgr} \quad \text{where} \quad \varepsilon'_{ijgr} \equiv \eta_{ijg} - \beta e_{ijg-r} \tag{11.3}$$

where q_{ijg} and q_{ijg-r} are positively correlated, and jointly normally distributed under the assumption that a_{ij}, e_{ijg-r} and η_{ijg} are each normally distributed.

In seeking to estimate the assumed underlying educational production function (11.1) through the regression equation (11.3) between pupil educational attainment at stage g and pupil prior attainment at stage g-r, both OLS and multilevel estimation techniques encounter the problem that the disturbance term ε'_{ijgr} in (11.3) is now negatively correlated with the explanatory variable of the observed pupil prior attainment variable q_{ijg-r}, since under the above assumptions:

$$\text{cov}(q_{ijg-r}, \varepsilon'_{ijgr}) = -\beta\sigma_e^2 < 0 \tag{11.4}$$

.

As in Gujarati (1995, p. 469-79), an OLS regression estimate of the coefficient β of the assumed underlying educational production function (11.1) would be not only biased but also *inconsistent* i.e. biased away from its true value even if the sample size increases indefinitely, with an asymptotic value to the estimate $\hat{\beta}$ of β given by:

$$\beta' \equiv \text{plim}\hat{\beta} = \beta/(1+(\sigma_e^2/\sigma_a^2)) = \beta(1-(\sigma_e^2/\sigma_{q_{g-r}}^2)) < \beta \qquad (11.5)$$

where σ_a^2 is the variance of the ability variable a_{ij} and $\sigma_{q_{g-r}}^2$ is the variance of the pupil prior attainment variable q_{ijg-r}. The effect of such a downward bias in the estimate of β is to pivot downwards the regression line, such as from the line AB with slope β to a line such as CD in Figure 11.1 below, around the mean point, such as at M.

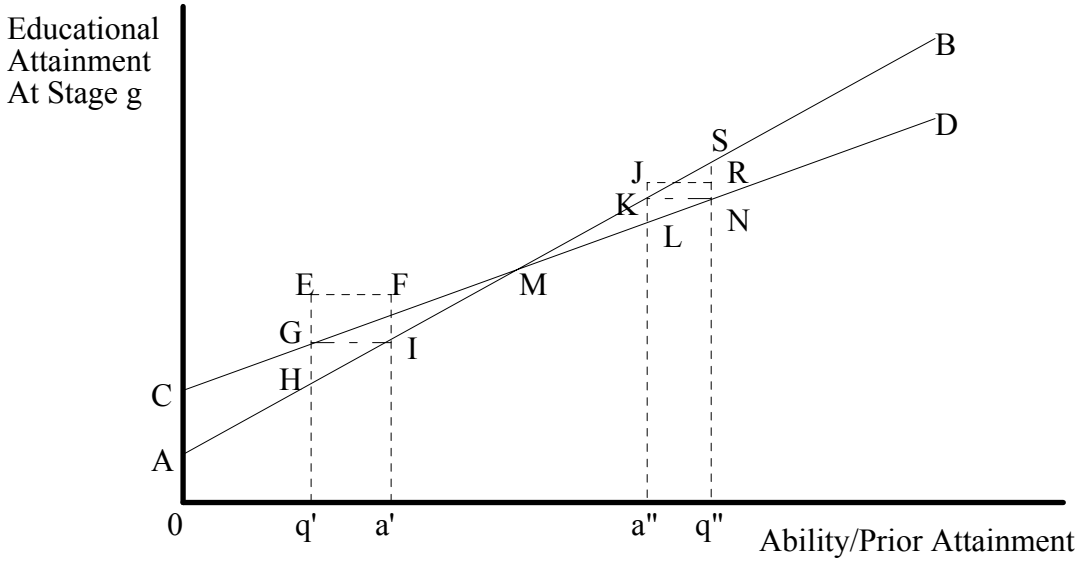

FIGURE 11.1

Such a pivoting of the estimated regression line risks biasing the estimation of pupil value added, if this is to be based upon underlying pupil ability. In particular, it will tend to reduce the estimated value added of those pupils, such as at E, whose prior attainment is below the average level of prior attainment (as tends to be the case in Academies). The estimated value added for a pupil at point E in Figure 11.1 will be reduced from EH to EG, based upon the estimated regression line CD, rather than the true underlying regression line AB corresponding to the educational production function (11.1). In contrast, the estimated value added of pupils with prior attainment levels greater than the average for all pupils in the sample will tend to be over-estimated. For a pupil at point R in Figure 11.1, the estimated pupil value added will be the positive amount RN, based upon the estimated regression line CD, rather than the negative amount SR, based upon the line AB corresponding to the educational production function (11.1).

However, this assumes that we do have available an unbiased estimate of pupil ability to insert into the estimated regression line CD, to generate our predicted value of educational attainment at stage g, from which to compute pupil value added. If all we have are observations of pupil prior attainment, then we need to take account of the positive correlation which exists between the observed level of pupil prior attainment and the error term e_{ijg-r} in (11.2) between pupil prior attainment and pupil ability. Such a positive correlation means that when pupil prior attainment is less than its mean level, as at point E in Figure 11.1, the observed value of pupil prior attainment, such as at q', tends to be lower than the level of prior attainment, a', that would correspond to the pupil's underlying ability. Conversely, when pupil prior attainment is greater than its mean level, as at point R in Figure 11.1, the observed value of pupil prior attainment, such as at q", tends to be greater than the level of prior attainment, a", that would correspond to the pupil's underlying ability. Adjusting for such bias in the estimates of underlying ability would mean making use of the true ability level a' for the pupil at point E in Figure 11.1. When combined with the true underlying regression line AB corresponding to the educational production function (11.1), this would lead to a pupil value added of IF that is equal to the lower estimate of EG that follows from the regression line CD based upon the observed level of prior attainment q'. Similarly, when the true level of ability a" for the pupil at point R in Figure 11.1 is combined with the true underlying regression line AB, this leads to a pupil value added of JK that is equal to the lower estimate of RN that follows from the regression line CD based upon the observed level of prior attainment q".

More generally, the expected value of pupil attainment at stage g, given the observed level of pupil prior attainment at stage g-r, equals:

$$E(q_{ijg} \left| q_{ijg-r}) = \alpha + \beta\mu_{g-r} - \beta(1-(\sigma_e^2 / \sigma_{q_{g-r}}^2))(q_{ijg-r} - \mu_{g-r}) = \alpha' + \beta' q_{ijg-r} \right. \qquad (11.6)$$

using Mood and Graybill (1963, p. 202), where $\alpha' \equiv \alpha + \beta(\sigma_e^2 / \sigma_{q_{g-r}}^2)\mu_{g-r}$. α' and β' are precisely the parameters of the regression line CD given by (11.5) and its associated interscept. The estimated regression line CD based upon pupil prior attainment thus gives an unbiased estimate of the expected value of pupil attainment at stage g, given the observed level of pupil prior attainment at stage g-r, and hence of individual pupil value added. Thus, even in the presence of measurement error, if we are forming our expectations of pupil attainment at stage

g based upon an observed variable, such as pupil prior attainment, OLS regression analysis will still give an unbiased estimate, and to this extent will still give an unbiased estimate of pupil value added by the school at stage g of the educational process. Moreover, the OLS estimate will be the best (i.e. minimum variance) linear unbiased predictor of the dependent variable conditional on the observed values of the explanatory variable, so long as the above joint normality assumption holds (see Fuller, 1987, p. 75).

b. The relevance of pupil prior attainment

It is important to consider here the underlying reasons for the disturbance term e_{ijg-r} in (11.2) in the deviation of the observed level of pupil prior attainment from pupil ability. Rather than being simply an error term, as our earlier interpretation supposed, the disturbance term e_{ijg-r} may convey important information about the level of motivation and effort that the individual pupil put into their studies by stage g-r of their education, compared to the mean level of such motivation and effort for all such pupils. This initial relative motivation and effort may indeed be relevant to the correct specification of the educational production function and the associated identification of the extent of the school's contribution at stage g of the educational process. It may then be more appropriate to write the educational production function (11.1) in the form:

$$q_{ijg} = \alpha + \beta a_{ij} + \gamma e_{ijg-r} + \eta'_{ijg} \quad \text{where} \quad \beta, \gamma > 0 \tag{11.7}$$

where η'_{ijg} is the new measure of school effectiveness for pupil i at stage g. If the normalisation of the pupil ability index a_{ij} gives it equal weight to the pupil motivation and effort variable e_{ijg-r} in the production of pupil prior attainment in (11.2) above, it is arguable that they should have equal weight in defining the initial starting point of the pupil for defining the contribution that is subsequently made by the pupil. In such as case, we may re-write (11.7) as:

$$q_{ijg} = \alpha + \beta(a_{ij} + e_{ijg-r}) + \eta'_{ijg} = \alpha + \beta q_{ijg-r} + \eta'_{ijg} \tag{11.8}$$

so that pupil prior attainment, rather than simply pupil ability, does become the correct explanatory variable for measuring pupil value added if relative pupil motivation and effort are relevant in this way. It will also be the correct explanatory variable if pupil prior attainment in addition adequately reflects the pupil's *initial accumulated stock of knowledge and skills*, that may well also be relevant for defining the educational baseline for stage g. Such an accumulated stock of knowledge and skills corresponds to the *intellectual capital* input into the educational process outlined by Hargreaves (2001). While pupil motivation and effort at stage g may be important factors in determining pupil attainment at stage g, the influence of any change in their level compared to their initial value, will be attributed here to the influence of the school during stage g of the educational process. There is then no measurement error in using the pupil prior attainment variable directly as the explanatory variable, so long as there are no further complicating factors.

An additional case of interest is where there are other factors that mean that underlying pupil ability may make a greater contribution than the initial level of pupil effort and motivation, in influencing educational attainment at stage g of the educational process. One such case is where the pupil has English as an Additional Language (EAL) and their underlying ability is initially impaired by their knowledge of English in securing a higher level of prior attainment at stage g-r, for any given relative level of motivation and effort. The reduction in this initial handicap as the pupil's knowledge of English increases over time will enable their underlying ability to make a greater contribution at stage g than it did at stage g-r, for a given relative level of motivation and effort. Since this additional contribution is strongly correlated here with their EAL status, the insertion of the additional EAL variable into the regression equation (11.8) can seek to correct for this factor.

Pupil prior attainment may reflect not just pupil ability and relative motivation and effort, but also wider socio-economic background factors which may limit the extent to which pupil ability is allowed to develop to its full potential. The fact that pupil prior attainment can encapsulate the combined effect of these many different influences itself reinforces the desirability of using it as a primary explanatory variable in the regression analysis to assess the value added by the school. However, if there are external factors, such as a change in the socio-economic circumstances of the pupil and their family, which change the balance between the different underlying variable that influence educational attainment between stage g-r to stage g

of the educational process, then such changes also need to be incorporated into the regression analysis to assess the value added by the school.

c. Measurement error in pupil attainment

Measurement error in the dependent error of pupil attainment at stage g will not itself bias the OLS estimate of the parameter β in (11.8). However, it will increase its standard error, which will be an increasing function of the variance of any measurement error for q_{ijg} (see Gujarati, 1995, p. 468). However, according to Davidson and McKinnon (2004, p. 313), "Unless the increase is substantial, this is not a serious problem".

Any remaining measurement errors in pupil prior attainment, or in other explanatory variables, will produce biased and inconsistent estimates of the coefficients of the underlying structural equation of the *educational production function*. As Bound *et al* (2001) note, "with measurement errors in more than one explanatory variable, the bias on any particular coefficient will involve multiple terms, and is hard to characterize. What should be clear is that without some knowledge of the distribution of the error…, the situation is hopeless – the data put no restrictions on [its] possible values" (ibid, p. 3716). As Goldstein (1995) notes more generally, "The topic of measurement error estimation is a complex one, and there are, in general, no simple solutions, except where the assumption of independence of errors on repeated measuring can be made. The common procedure, especially in education, of using 'internal' measures based upon correlation patterns of test or scale items, is unsatisfactory for a number of reasons and may often result in reliability estimates which are too high". As Browne *et al* (2001, p. 4) note: "in the case where there are errors of measurement in a predictor that has a random coefficient, likelihood and moment based techniques become intractable". Alternative approaches to estimating the effect of measurement error in a predictor are the use of Markov Chain Monte Carlo (MCMC) estimation techniques if the error variance, as discussed in Browne *et al* (2001), and the use of bootstrap procedures discussed in Hutchison *et al* (2003).

As Bound *et al* (2001, p. 3709) emphasise, "Standard methods for correcting measurement error bias, such as instrumental variables estimation, are valid when errors are classical [i.e.

independent of the true level of the variable and of all the other variables, of the measurement error in other variables, and of the stochastic disturbance term in the model] and the underlying model is linear, but not, in general, otherwise …. Not only can standard fixes not solve the underlying problem, they can make things worse!". In order to obtain greater knowledge of the nature of the error terms, Bound *et al* (2001) advocate the use of *validation studies* to compare observed data with data obtained under circumstances that are less prone to measurement error. They note (*ibid,* p. 3709) that: "One general conclusion from the available validation evidence is that the possibility of non-classical measurement error should be taken much more seriously by those who analyze survey data, both in assessing the likely biases in analyses that take no account of measurement error and in devising procedures that 'correct' for such error".

Gujarati (1995, p. 470) concludes that: "There is really no satisfactory answer to the measurement errors problem. That is why it is so crucial to measure the data as accurately as possible". However, as Newton (2005, p. 436), of the Research and Statistics Team, Qualifications and Curriculum Authority, notes: "If we accept the need for educational measurement then we must accept the inevitability of error. There is no such thing as perfection when it comes to validity, reliability or comparability". Some degree of bias in the estimated coefficients in a value-added analysis away from their true value in an underlying educational production function will then be a consequence of measurement errors in the explanatory variables. However, knowledge of the true values in an underlying educational production function becomes less critical when value added analysis is itself defined in terms of a comparison between achieved levels of pupil attainment and their *predicted* levels, conditional on the *observed* values of the explanatory variables.

There is nevertheless a need for further research using Monte Carlo simulation into the **sensitivity** of estimates of pupil- and school-level value added, such as those provided by the Contextual Value Added models discussed in Section 8, to possible variations in the observed data within the range of their likely inaccuracies. This is particularly the case when these estimates are obtained by an iterative process, such as that used by multilevel modelling, where parameter stability and convergence to a locally close parameter estimate may not be guaranteed under such variations, and when the number of explanatory variables and associated coefficients is large. As noted by Kreft *et al* (1994, p. 334) in their review of software packages for estimating multilevel models (including those which use Maximum Likelihood estimation): "In general, it follows from our analysis that even if we restrict

ourselves to only two-level models with random slopes, we have very complicated likelihood surfaces. Maximising the likelihood is inherently a difficult problem, unless the model is approximately true and the sample size is really large (in which case OLS will give very good starting values). Investigators (if the past is any indication) will tend to choose models that are too complicated … This leads to impossibly difficult search problems over the space of models and to impossibly difficult likelihood maximisation problems. None of the programs reviewed here can handle such problems gracefully".

While the inclusion of a large number of explanatory variables, as in the Contextual Value Added models discussed in Section 8 above, has its own attractions, greater **parsimony** in the selection of variables may increase the **robustness** of the resultant estimates to departures from the assumptions of the underlying multilevel model. The need for further research into the impact of such departures, and the relative merits of multilevel and simpler estimation techniques in the face of such departures, is emphasised also by de Leeuw and Kreft (1995) and Morris (1995). As Goldstein (1991, pp. 90-91) has noted: "as in all statistical models, the estimates we obtain are sensitive to the assumptions we make, and this will tend to be more important for residual estimates than for estimates of the fixed and random parameters in the model … As with many new techniques that promise a substantial advance in understanding, multilevel modelling is not a panacea. Its power is limited, and it is certainly not a magic wand that will allow us automatically to make definitive pronouncements about differences between individual schools".

d. Endogeneity bias

As noted above, a second way in which there may be a breach of the assumption that the explanatory variables that enter into the educational production function are **uncorrelated** with the residuals in the regression analysis for the school's educational performance is if there are **other important inter-relationships** between these explanatory variables and educational performance in addition to that described by the concept of the educational production function. As in Mayston (2000), these include other important inter-relationships may include:

i. the school's level of educational performance influencing pupil numbers through parental **demand for places** at the school, with pupil numbers entering into the educational production

function if there are fixed costs and **economies of scale** in the production of educational output for the school (see e.g. Bradley and Taylor, 1998);

ii. the school's level of educational performance influencing **local house prices** and hence the level of income of parents who can afford to live in the local area, together with a correlation between **parental income** and the characteristics of the school's pupil intake;

iii. the school's level of educational performance influencing the **quality of teaching staff** which the school it is able to attract;

iv. the school's level of educational performance influencing the **school's income and resources** which it has available to it.

In addition, a correlation can occur if:

v. an **intermediate level of pupil performance**, such as at KS3, is used as a **prior attainment** variable to explain pupil performance at a later stage, such as at GCSE, and the school effect component of the residual in a value-added analysis for this phase of the educational process is correlated with the school effectiveness at the earlier phase of the educational process **within the same school**, such as from KS2 to KS3.

vi. individual pupil performance at an intermediate stage, such as at KS3, is correlated with **additional unmeasured factors**, such as pupil motivation, that influence the pupil-level residuals at the later phase of educational progress, such as from KS3 to GCSE.

The strength of the correlation involved in i**. – iv.** above may be reduced by:

I. time lags in the impact of the school's level of educational performance on pupil demand, local house prices, pupil characteristics, teacher characteristics and school income; and

II. a value-added analysis that generates **value-added residuals** to which parents, teachers and school income are less sensitive than they are to **school league table information** on the absolute level of examination results for the school.

The existence of factors v. and vi. above would in particular suggest that the estimates of both an OLS regression analysis and multilevel modelling may be **biased** by the existence of such endogeneity if they focus on value added from KS3 to GCSE. Where both GCSE performance and KS3 performance take place within the same school, the endogeneity bias is likely to be significantly reduced by focusing not on pupil progress from KS3 to GCSE, but rather upon pupil progress from KS2 to GCSE, and possibly also on pupil progress from KS2 to KS3, where the prior attainment variable at KS2 is less likely to be correlated with the secondary school residuals.

In cases where endogeneity remains a problem, Instrumental Variables techniques may produce unbiased estimates if appropriate instruments are available (see Mayston, 2002). For cases where the correlation is not at the lowest hierarchical level, here the pupil level, Rice *et al* (1999) propose use of a conditioned version of the multilevel modelling estimation procedure of Iterative Generalised Least Squares (IGLS). Using datasets of pupil progress from GCSE to A-level, Spencer and Fielding (2002) show that the use of Bayesian Inference Using Gibbs Sampling (BUGS) techniques can produce parameter estimates with lower standard errors to those produced by instrumental variables methods for tackling the endogeneity problem.

The existence of possible endogeneity bias is one of several issues in the estimation of pupil value added that merit further detailed investigation within a follow-up research project that is not restricted simply to the evaluation of the Academies programme. Other issues that merit such an examination include the stability of the parameter estimates produced by multilevel estimation, the role of resource variables, and the comparative performance of models with different degrees of parsimony in their choice of explanatory variables. Each of these topics raises important and interesting questions concerning the precise application of value added estimation that apply much more generally than the evaluation of the Academies programme, and deserve a fuller investigation in their own right. The DfES (2005) Contextual Value Added has nevertheless made considerable progress in the development of the estimation of pupil value added. While any conclusions based upon it will be contingent upon the assumptions implicit in it, this will be true of any existing state of knowledge and should not prevent provisional conclusions based upon it being made, albeit subject to these caveats.

12. PROGRAMME EVALUATION AND COMPARISON GROUPS

a. Estimating the programme impact

One approach to incorporating the evaluation of the impact of the Academies programme into a value-added framework is through the use of a dummy variable $d_{jt} = 1$ to designate the possession of Academy status by school j at time t and $d_{jt} = 0$ to indicate the contrary. The multilevel equation at time t would now become:

$$y_{ijt} = \alpha_t + \sum_{k \varepsilon P} \beta_{kt} x_{kijt} + \sum_{h \varepsilon S} \gamma_{ht} z_{hjt} + d_{jt} \delta_t + \theta_{jt} + \varepsilon_{ijt} \qquad (12.1)$$

where y_{ijt} is the educational outcome score for pupil i in school j at time t, x_{kijt} is the value of the kth pupil-level variable for pupil i in school j at time t, z_{hjt} is the value of the hth school-level variable for school j at time t, α_t, β_{kt} , γ_{ht} and δ_t are fixed parameters, and θ_{jt} and ε_{ijt} are assumed to be stochastic variables at the school and pupil levels respectively. The set P of pupil-level variables will include pupil prior attainment scores and other relevant pupil-level variables, such as gender.

Regression-based estimation procedures, such as OLS and those conventionally used in multilevel modelling, assume that the explanatory variables, including the programme participation variable d_{jt} in (12.1), are uncorrelated with the stochastic disturbance terms. One method of ensuring this lack of correlation would be through an **experimental design** (see Fitz-Gibbon and Morris, 1987) which ensured a random allocation of schools and pupils to the programme to be evaluated, here the Academies programme. This approach would seek to replicate the advantages of the use of **randomised control trials** (RCTs) in health sciences and elsewhere (see e.g. Chambers *et al*, 1981; Montgomery *et al*, 2004).

Such a random selection from a wider population of schools and their associated pupils would ensure that the comparison group of schools and pupils who were not part of the programme were statistically equivalent to the participating group in all relevant variables except their participation in the programme. The experience of the comparison group would then form a **counterfactual** set of outcomes to the experience of the participating schools and pupils, with the only relevant difference between the two groups being participation in the programme. The

impact, δ_t, of positive participation in the programme at time t could then be estimated through comparing the mean level of the outcomes for the participating schools with the mean level for the comparison group schools, i.e.

$$\hat{\delta}_t = Y_t^A - Y_t^C \tag{12.2}$$

where ^ designates an estimated value and Y_t^A and Y_t^C are the mean values of the outcomes y_{ijt} for pupils in participating and comparison group schools respectively (c.f Blundell and Costa Dias, 2000).

In the absence of a formal experimental design, the extent to which selection into the programme is random may to some extent be gauged by examining the similarity of the distributions of the observable characteristics, x_{kijt} and z_{hjt}, across the participating and comparison group schools. At least as far as the observable characteristics are concerned, a comparison group might indeed be chosen using the criterion of **similarity in the distribution** of these characteristics with that across the participating schools in order to approximate the outcome of a random selection of schools into the programme.

However, even if the observable characteristics do have a similar distribution, there may still still remain differences in the unobservable terms that influences the school- and pupil-level effects in (12.1) and which cause their expected values across participating and comparison group schools to differ. In such a case, the estimator (12.2) will not provide a consistent (i.e. asymptotically unbiased) estimate of the underlying impact parameter δ_t for the programme. Such will be the case if these unobservable effects influence the decision of whether or not a school participates in the Academies programme and whether or not particular pupils attend an Academy school.

In the case of the Academies programme, the availability of the Pupil Level Annual School Census (PLASC) database means in principle that **repeated cross-section value-added analyses** can be carried out on data for years **before and after** the implementation of the programme using a **'difference-in-differences'** (diff-in-diffs) approach.

The diff-in-diffs estimator of the impact of the programme is given by:

$$\hat{\delta}_{tD} = (Y_t^A - Y_o^A) - (Y_t^C - Y_o^C) \tag{12.3}$$

and involves the changes in mean outcome levels for the Academies schools and the control group schools at year t after the programme has been implemented compared to before the start of the Academies programme in a base year 0. The diff-in-diffs estimator (12.3) will provide a **consistent estimator of the programme impact** δ_t in (12.1) so long as:

$$\begin{aligned} \Delta_t^A &\equiv E(\theta_{jt} + \varepsilon_{ijt} \left| d_{jt} = 1\right.) - E(\theta_{jo} + \varepsilon_{ijo} \left| d_{jt} = 1\right.) \\ &= \Delta_t^C \equiv E(\theta_{jt} + \varepsilon_{ijt} \left| d_{jt} = 0\right.) - E(\theta_{jo} + \varepsilon_{ijo} \left| d_{jt} = 0\right.) \end{aligned} \tag{12.4}$$

i.e. the changes in the expected values of the stochastic school-and pupil-level effects over the period are the same for Academies and comparison group schools, together with:

$$\begin{aligned} &E(x_{kijs} \left| d_{jt} = 1\right.) = E(x_{kijs} \left| d_{jt} = 0\right.) \text{ for all k}\varepsilon\text{P and for s= 0, t} \\ &E(z_{hjs} \left| d_{jt} = 1\right.) = E(z_{hjs} \left| d_{jt} = 0\right.) \text{ for all h}\varepsilon\text{S and for s= 0, t} \end{aligned} \tag{12.5}$$

i.e. the same mean values to the pupil- and school-level variables in the participating and the comparison group schools.

Under these conditions, a 'more robust estimate of the impact' (Blundell and Costa Dias, 2000, p. 437) of the programme can be made than is possible through using either Instrumental Variables (IV) estimation or the two-step Heckman selection estimator (Heckman, 1979) to model the participation decision when only one cross-section of data is available. The use of the Instrumental Variables technique itself depends upon being able to find a suitable instrument which determines programme participation, but which is not itself determined by the factors which affect the outcomes (see Bryson *et al*, 2002). In addition the estimates from the two-step Heckman selection estimation technique can be very sensitive to the assumptions it makes regarding the distribution of the unobserved variables (*ibid*, p. 10; Puhani, 2000),

though with Vella (1998) examining extensions of the Heckman approach to relax these distributional assumptions and its parametric assumptions.

If conditions (12.4) and (12.5) hold, the diff-in-diffs estimator (12.3) will also provide a consistent estimator of the **average impact** δ_t' of the Academies programme on the participating schools, even where the impact of the programme is **not uniform** across all participating schools (c.f. Blundell and Costa Dias, 2000, p. 442). In contrast to the case of a **homogeneous programme impact**, δ_t , in (12.1), we may have instead a **heterogeneous impact** δ_{jt} of the Academies programme that **differs across individual participating schools** j. (12.1) can then be modified to:

$$y_{ijt} = \alpha_t + \sum_{k \varepsilon P} \beta_{kt} x_{kijt} + \sum_{h \varepsilon S} \gamma_{ht} z_{hjt} + d_{jt} \delta_{jt} + \theta_{jt} + \varepsilon_{ijt} \qquad (12.6)$$

with δ_t' the expected value of δ_{jt} for participating schools.

We may relax the condition (12.5) through use of the **value-added adjusted difference-in-differences estimator**:

$$\hat{\delta}_{\text{tDV}} = (V_t^A - V_o^A) - (V_t^C - V_o^C) \text{ where } V_s \equiv E'(V_{ijs}) \;\&\; V_{ijs} \equiv y_{ijs} - \alpha_s - \sum_{k \varepsilon P} \beta_{ks} x_{kijs} - \sum_{h \varepsilon S} \gamma_{hs} z_{hjs} \qquad (12.7)$$

for s = 0, t, and where E' denotes the mean value across pupils and schools in the relevant group and the superscripts A and C refer to the Academies group and the comparison group respectively.

Under condition (12.4), (12.7) will now yield a consistent estimator of the programme impact δ_t in the homogeneous case (12.1) or of the average programme impact δ_t' on the participant schools in the heterogeneous case (12.6). In the heterogeneous case, the average programme impact δ_t' on the participant schools corresponds to **'the Effect of Treatment on the Treated' (TT).** In the heterogeneous case, TT must be distinguished from **'the Average Treatment Effect' (ATE)** for a school chosen at random from the population who would be eligible for the programme and from **'the Marginal Treatment Effect' (MTE)** that corresponds to the average effect for potential participants in the programme who are on the margin of

indifference of whether or not they participate (see Aakvik, Heckman and Vytlacil, 2005, p. 20). While the diff-in-diffs estimator succeeds in relaxing the assumption that selection into the programme only depends on the observable variables, this means that unobserved components of the programme impact may still affect participation in the programme as temporary individual-specific effects (see Blundell and Costa Dias, 2000, p. 442), so that if treatment effects are heterogeneous across programme participants, the effect of treatment on the treated may differ from the average treatment effect for the wider population of schools who might have participated in the programme.

Angrist (2004) establishes conditions under which estimates of the Average Treatment Effect (ATE) can still be derived from estimates of the Local Average Treatment Effect (LATE) using instrumental variables, even if treatment effects are heterogeneous across programme participants. In particular, he considers the case where participation in the programme is determined by a criterion of the form:

$$D_i = 1 \text{ when } \gamma_o + \gamma_1 Z_i > \eta_i \tag{12.8}$$

where Z_i is a (0,1) binary instrument and η_i is a random error term that is independent of Z_i, with:

$$D_i = D_{oi}(1 - Z_i) + D_{1i} Z_i \tag{12.9}$$

so that D_{oi} indicates whether of not individual i would be a programme participant if $Z_i = 0$ and D_{1i} indicates whether of not individual i would be a programme participant if $Z_i = 1$.

When the impact of programme participation (i.e. the treatment effect) is heterogeneous across individuals and individuals themselves have different characteristics, the expected outcomes Y_{1i} and Y_{0i} from programme participation and non-participation respectively may vary according to who is selected into the programme and who is not selected into the programme. Angrist (2004) therefore models **conditional expectation functions** (CEFs) for the expected outcomes Y_{1i} and Y_{0i} in the form:

$$E(Y_{1i} \,|\, D_{oi}, D_{1i}) = a_1 + b_{10} D_{0i} + b_{11} D_{1i} \; ; \quad E(Y_{0i} \,|\, D_{oi}, D_{1i}) = a_0 + b_{00} D_{0i} + b_{01} D_{1i} \tag{12.10}$$

The first restriction that is sufficient to ensure that the Average Treatment Effect (ATE) can be identified from the Local Average Treatment Effect (LATE) that is estimated by instrumental variable analysis is that:

$$b_{00} = b_{01} = b_{10} = b_{11} = 0 \tag{12.11}$$

implying that there is **no selection bias** for participation in the programme. Participants in the programme are a representative sample of all individuals in the population at large, so that knowledge of their likelihood of selection does not influence the expectations of the outcomes Y_{1i} and Y_{0i}. We then have LATE = $a_1 - a_0$ = ATE. However, in the case of Academies, the schools selected in the programme are clearly not a representative sample of all schools in the wider population, though they may be of a narrower subset of such schools.

A second restriction which is sufficient to ensure LATE = $a_1 - a_0$ = ATE is that:

$$b_{00} = b_{10};\ b_{01} = b_{11} \tag{12.12}$$

so that the value of D_{0i} has the same effect on the expected values of both Y_{1i} and Y_{0i}, and hence does not affect the difference in these expected values, and similarly for D_{1i}.

A third condition under which the Average Treatment Effect (ATE) can be estimated from LATE is when:

$$b_{00} = \theta b_{01}; b_{10} = \theta b_{11} \qquad \text{for} \quad \infty > \theta > 0 \tag{12.13}$$

so that the individual values of D_{0i} and D_{1i} affect the expected values of both Y_{1i} and Y_{0i} only via the composite index $D_{0i} + \theta D_{1i}$.

One special case of a heterogeneous programme impact that is of interest here is where the programme impact varies according to how long a school has been open as an Academy. If the programme impact is proportional to the number of years that any given school has been open

as an Academy, a consistent estimate of the impact, $\hat{\delta}^{a}_{tDV}$, of the programme per year of participation in the programme may be obtained in a parallel way to (12.7) as:

$$\hat{\delta}^{a}_{tDV} = [(V^{A}_{t} - V^{A}_{o}) - (V^{C}_{t} - V^{C}_{0})][n_{A(t)} / \sum_{j \varepsilon A(t)} T_{j}] \qquad (12.14)$$

where T_j is the number of years which school j has been open as an Academy, *A(t)* is the set of all Academies that have been opened by time t and $n_{A(t)}$ is the number of Academies which have been opened by time t.

The remaining condition (12.4) will be satisfied if

$$E(\theta_{jt} + \varepsilon_{ijt} \,|\, d_{jt} = 1) = E(\theta_{jo} + \varepsilon_{ij0} \,|\, d_{jt} = 1) \;\; \& \;\; E(\theta_{jt} + \varepsilon_{ijt} \,|\, d_{jt} = 0) = E(\theta_{jo} + \varepsilon_{ijo} \,|\, d_{jt} = 0) \qquad (12.15)$$

so that the **expected values** of the random individual pupil- and school-effects for the programme and comparison group do not change over time. (12.1) and (12.7) still permit a **common overall rate of underlying school improvement**, for both the Academies school and the comparison group schools in the absence of the impact of the Academies programme, through α_t differing from α_o. In addition, (12.15) is consistent with individual school-specific effects, if these do not change over time. This may include **heterogeneity** in the production possibilities that individual schools face, of the kind emphasised in the stochastic frontier literature discussed in Section 9 above, so long as these effects are time-invariant over the period of the evaluation.

Some persistence in the rate of improvement or deterioration of individual school effects over time is also consistent with (12.4), so long as there is a common overall rate of change in the **expected value** across individual schools of the sum of the pupil and school effects in the Academies and comparison groups. If there is an underlying **differential overall rate of improvement** between the two groups that is additional to the impact of participation in the programme, the differentially-adjusted estimator of Bell, Blundell and Van Reenan (1999) based upon repeated comparisons of the RHS of (12.7) over several years can still yield a consistent estimator of the programme impact. Such an adjustment would be desirable if there is evidence of a **selection effect** for schools to participate in the Academies programme in

favour of schools who would otherwise have a different rate of improvement than the schools in the comparison group.

Another selection effect of the Academies programme may be to attract more able pupils to the Academies who would otherwise not attend the schools which became Academies. Where this involves differences in the observable prior attainment scores and observable pupil characteristics within the set P of such pupil characteristics in (12.1) or (12.6), systematic adjustment is made in (12.7) for such differences. Where it involves differences in the mean value of the unobservable pupil-effects ε_{ijs} across the Academies and the comparison group schools that change differentially as the Academies programme is implemented, the condition (12.4) may not hold. The diff-in-diffs estimator (12.7) will then not give a consistent estimator of the pure programme impact δ_t' , but instead will include an effect due to the change in the mean value of the unobservable pupil-effects ε_{ijs} across the Academies and the comparison group schools. It might be argued that this will provide an estimate of the overall impact of the Academies programme, including that due to the **increased ability** of the schools participating in the Academies programme to attract more motivated pupils.

One method of seeking to eliminate this secondary effect from the estimate of δ_t' would be to estimate the value added in (12.7) for only those pupils who were **originally in the predecessor schools** to the Academy schools, and hence who did not self-select into the Academies sample as a result of the school's participation in the Academies programme. However, this would still include pupils who might have continued to attend the school as a result of its Academy status but who would otherwise have moved elsewhere. Some indication of the importance of these effects may be gained through examining the **changes in the profile of prior attainment and other pupil-level characteristics**, and rate of **pupil mobility**, of Academy schools compared to those in the comparison group over the evaluation period. Where there are measurable differences, they can be included within the sets P and S of relevant pupil- and school-level variables, with only significant changes in unmeasurable influences causing a breach of condition (12.4).

b. Matching

A further approach to the evaluation of the programme impact is through the use of **matching** procedures to pair each participant in the programme with a corresponding member of the comparison group that does not participate in the programme. The **conditional independence assumption** that is made in much of the literature on matching involves the assumption that for the same set of observable characteristics X, the outcomes, here for $\{y_{ijt}\}$, are the same for the comparison group as they would have been for programme participants in the absence of the programme. Unobservable characteristics are assumed to play no part in distinguishing programme participants from non-participants in the comparison group. This means that knowledge of the observable characteristics X for participants in the programme and of the outcomes for the corresponding members of the comparison group with the same value of the vector X is sufficient to construct the counterfactual outcome for the programme participants had they not taken part in the programme. The impact of the programme for each value of X can then be evaluated as the difference between the mean value of the outcome for participants in the programme with this value of X and matching members of the comparison group with the same value of X. In the context of eqns (12.1), (12.6) and (12.7), the observable characteristics for any given school j at time t correspond to the set $X_{jt} = \{x_{kijt}, z_{hjt} \mid k\varepsilon P, h\varepsilon S \,\&\, i\varepsilon I_j\}$, where I_j is the relevant set of pupils in school j.

Such matching according to the observable characteristics X can be preferable to a random selection of members of the comparison group from a wider population of schools and pupils. This is because it can increase the likelihood that members of the comparison group could have been chosen for participation in the programme and bring closer together the expected value of the unobservable characteristics of those in the comparison who would and would not have been eligible to participate in the programme (see Blundell and Costa Dias, 2000, p. 447). This makes it more likely that the conditional independence assumption will actually hold.

However, pairwise matching according to the vector X is likely to be difficult to achieve in practice, once X involves a substantial number of variables, each of which may take on a large number of values. In order to overcome this problem, Rosenbaum and Rubin (1983) suggest use of **propensity score matching** that seeks to match programme participants with a corresponding comparison group member not with the same X value but simply with the same

value of the probability, p, that they would have participated in the programme. By virtue of the conditional independence assumption, p is simply a function, p(X), of the observable characteristics X, and provides a single scalar variable on which matching is required under propensity score matching, rather than on the entire multi-dimensional vector X. However, the use of propensity score matching, in common with most other methods of matching, requires the existence of a '**common support**' for X, i.e. the same set of X values within the comparison group as in the group of programme participants. If this condition does not hold initially, some of the observations in the programme group may need to be discarded until a matching with the available comparison group is achieved.

Hahn (1998) shows that, if the propensity score is known, its use can reduce the asymptotic variance for the estimate of the average treatment effect on the treated, though conditioning on the propensity score is not necessary for this effect to be efficiently estimated. Heckman *et al* (1998a) show that, if exclusion restrictions are placed upon the set of variables from X that are used to estimate the propensity score, the use of propensity score matching does not necessarily reduce the variance of the resulting estimate, even when the propensity score is known. When it is unknown, the estimation of the propensity score and the process of matching both generate additional sources of variation. Heckman *et al* (1998a, p. 281) also show that exclusion restrictions which reduce the dimensionality of the set of variables that are used to estimate the propensity score help to reduce the asymptotic variance of the matching estimator by reducing the estimation error from the estimation of the propensity score. Exclusion restrictions which reduce the number of variables which are used to determine outcomes also reduce the asymptotic variance of the matching estimator.

Smith and Todd (2005a) found that estimates based on the use of propensity score matching, of the impact of the US National Supported Work programme that was previously studied by Dehejia and Wahba (2002), are "highly sensitive to both the set of variables included in the scores and the particular analysis sample used in the estimation" (*ibid*, p. 305). In addition, Smith and Todd (2005b) emphasise the sensitivity of the estimates of programme impact to small sample sizes. Similarly, Dehejia (2005) finds that the estimated treatment effect that results from use of propensity score matching can be very sensitive to the specification of the function defining the propensity score, and concludes that: " Propensity score matching does not provide a silver-bullet, black-box technique that can estimate the treatment effect under all circumstances". In contrast, Smith and Todd (2005a) conclude that the **difference-in-**

differences matching estimators developed by Heckman *et al* (1997) and Heckman *et al* (1998b) "perform substantially better that the corresponding cross-sectional matching estimators" (*ibid*, p. 347), such as those associated with propensity score matching.

Heckman *et al* (1998a) prove that, rather than requiring the conditional independence assumption, what is required for use of the propensity score matching method is a weaker **'mean independence condition'**. This requires that the expected counterfactual outcome for programme participants had they not participated in the programme is the same as the expected outcome for non-participants **with the same propensity score** (rather than necessarily with the same complete set X of observable variables that may influence outcomes). However, in an evaluation of a job training programme, Heckman *et al* (1997) test and reject both the conditional independence assumption and the weaker mean independence condition. They also test, but do not reject, the even weaker identifying assumption for their **'conditional difference-in-differences'** estimator, which computes the 'diff-in-diffs' estimator (12.3) conditional on X. This assumption is the '**difference-in-differences mean independence**' condition that is equivalent here to:

$$E(\theta_{jt} + \varepsilon_{ijt} - \theta_{jo} - \varepsilon_{ijo} \,|\, X, d_{jt} = 1) = E(\theta_{jt} + \varepsilon_{ijt} - \theta_{jo} - \varepsilon_{ijo} \,|\, X, d_{jt} = 0) \qquad (12.16)$$

so that conditional on X, there is no difference in the expected values of the changes in the sum of pupil and school effects between the Academies and Comparison Groups. Where matching can be carried out on the probability of programme participation p, as in (12.24) below, the equivalent condition is

$$E(\theta_{jt} + \varepsilon_{ijt} - \theta_{jo} - \varepsilon_{ijo} \,|\, p, d_{jt} = 1) = E(\theta_{jt} + \varepsilon_{ijt} - \theta_{jo} - \varepsilon_{ijo} \,|\, p, d_{jt} = 0) \qquad (12.17)$$

In both cases, unobserved variables, in the form of individual school-specific effects, are again permitted to influence participation, so long as the school-specific effects are constant over time. A corresponding **'regression-adjusted conditional difference-in-differences' matching estimator**, where the regression coefficients are estimated from regression analysis across the Comparison Group, is found by Heckman *et al* (1997, p. 631) to be 'an effective method in reducing bias' in their study. They note that 'it is more demanding in terms of its data requirements than the cross-sectional matching estimators because it requires pre-programme

data'. Where there is reliable relevant data both before and after an Academy opened, the use of such an estimator thus becomes feasible.

If one assumes a specific functional form for the regression equation, such as (12.1), estimation of the relevant relationship between the observable variables and the outcomes for the treatment and comparison groups will yield predictions of the respective outcomes for programme participants with and without the programme. As Blundell and Costa Dias (2000, p. 449) note, "In this case, one can easily guarantee that outcomes being compared come from populations sharing exactly the same characteristics" and that "not even the common support requirement is needed to estimate the impact of treatment on the treated – a simple OLS regression using all information on the treated and non-treated will consistently identify" the average programme impact. However, if the regression analysis is not to involve **extrapolations** on the basis of a potentially inappropriate functional form outside the common area of the explanatory variables, the comparison group should be chosen to match as closely as possible the underlying characteristics of the programme participants.

A remaining potential source of selection bias under matching arises if the unobserved variables that influence the programme participation decision include transitory individual effects. This might arise if a school-level variable, such as the percentage of pupils eligible for FSM or average KS2 point score, were used in the participation decision, but were subject to **transitory disturbances**, such as measurement error. Matching on the basis of such participation variables as pre-test scores is then opposed by some authors, such as Kenny (1975) and Preece (1989), because of its dependence upon distortionary transitory mean-reverting effects. The transitory drop in average earnings of participants in government training programmes identified by Ashenfelter (1978), which yields an upward bias in the estimated programme impact, has been found to exist more widely by Heckman and Smith (1999). However, they also found that the use of local linear matching and of conditional difference-in-differences estimators substantially reduced, though did not eliminate, the extent of selection bias in their non-experimental estimates of the impact of participation in the training programmes.

c. Generating the Comparison Groups

According to DfES (2003c), "The Academies programme aims to challenge the culture of educational underattainment and to deliver real improvements in standards. All Academies are located in areas of disadvantage. They either replace one or more existing schools facing challenging circumstances or are established where there is a real need for additional school places...Academies will help break the cycle of underachievement in areas of social and economic deprivation whether in inner cities, suburban or rural areas".

An important indicator of **under-achievement and under-attainment** for the target pupil intake of Academies is that of low pupil performance at KS2. A school which has had an average KS2 score for its pupil intake that falls within the lower tail of the national distribution of the schools' average KS2 intake score has been facing educational under-attainment in its pupil intake. Defining this as a key selection variable will at the same time enable a range of values of other associated socio-economic variables to be included in the sample, in a way that reflects the range of challenging socio-economic circumstances that foster the under-attainment at which the Academies programme is aimed.

An alternative approach would be to focus on another single variables, such as the percentage of pupils in the school entitled to Free School Meals (FSM) or for whom English is an Additional Language, or on combinations of such variables. The OFSTED (2000) report *Improving City Schools*, for example, examined secondary schools which were 'more effective than others in similarly disadvantaged areas' and compared their performance with all secondary schools which had greater than 35 per cent of pupils entitled to FSM, as well as with all non-selective secondary schools. However, it noted that the FSM indicator was 'a meagre guide to the reality' (*ibid*, p.10) of the causes of disadvantage and under-attainment. The OFSTED (2003) report *Excellence in Cities and Educational Action Zones: Management and Impact* simply compared Excellence in Cities (EiC) and schools in Education Action Zones (EAZs) with the national average and with non-EiC schools. The NFER-LSE-IFS Evaluation of Excellence in Cities (Stoney *et al*, 2002) in contrast selected for the 296 EiC schools "33 comparison schools, chosen from non-EiC areas that were located in broadly comparable socio-economic circumstances".

Since the primary focus of the Academies programme is on under-attainment, the KS2 indicator has the advantage that it focuses on the major variable of the average level of **pupil prior attainment** of pupils on entering secondary education, with DfES (2005) noting that "prior attainment is by far the strongest predictor of outcomes". The prior attainment variable is itself the result of many local influences, including socio-economic deprivation, that influence the level of educational under-attainment of the pupil intake into the school. If under-attainment is the predominant criterion for participation in the Academies programme, one application of propensity score matching would be to infer that schools with the same level of average KS2 intake score faced the **same probability of being chosen for participation** in the Academies programme. An indicator based on the school average KS2 intake score can moreover be computed from the PLASC database. A Comparison Group for the Academies schools can be generated from schools with similar average KS2 scores in the base year of 2002 before the start of the Academies programme.

To provide a further benchmark against which the performance of the open Academies can be compared, additional variables may be included in the determination of the propensity score to yield a further Comparison Group for the open Academies. As noted above, propensity score matching enables several variables to be taken into account in estimating the determinants of the probability of a school being selected for participation in the Academies programme. These variables may include not only the average KS2 score of the pupil intake, but also additional variables, such as the percentage of pupils who were eligible for Free School Meals, the proportion of boys in the school population, the proportion of pupils assessed as having Special Educational Need, with and without statements, and ethnicity variables. These variables may form a subset Z of all the variables X that may influence outcomes, with the associated exclusion restrictions resulting in reduced asymptotic variance of the matching estimator (Heckman *et al* , 1998a, p. 281).

The estimation of the propensity score may then be achieved through use of **probit** or **logit** analysis (see Davidson and MacKinnon, 1993, pp. 514-5). Probit analysis results in the estimation of the propensity score for any given school h as the probability of its participation in the Academies programme given by:

$$p_h(X) = N(Z_h b) \qquad (12.18)$$

where N is the cumulative normal distribution function, Z_h is school h's vector of the values of its variables in the set Z, and b is a vector of coefficients reflecting the importance of each variable in the determination of its probability of participation in the programme. Logit analysis would compute a similar propensity score, given by:

$$p_h(X) = (1+ \exp(-Z_h b))^{-1} \qquad (12.19)$$

d. Matching estimators

Once each Comparison Group has been identified, the programme evaluation can make use of a 'regression-adjusted conditional difference-in-differences' matching estimator of the kind which, as noted above, Heckman *et al* (1997, p. 631) found to be an effective method for matching and reducing estimation bias due to selection. Such a matching estimator (*ibid*, pp. 629-31) can be expressed in the general form:

$$D_t(R) = \sum_{j \varepsilon A(t)} \omega_{n_o n_{A(t)}}(j)[(V_{jt} - V_{jo}) - \sum_{h \varepsilon C} W_{n_o n_{A(t)}}(j,h)(V_{ht} - V_{ho})] \quad for\ X \varepsilon R \qquad (12.20)$$

where R is a subset of the support of X for those values of X which prevail for the open Academies. (12.20) involves a weighted sum across these Academies of the change which each Academy has achieved over time in its value added, compared to a weighted sum of the changes in the value added which have been achieved by each school in the comparison group C. There are several ways in which the weights in these weighted sums may be chosen. One method would make use of a symmetric, nonnegative, unimodal '**kernel**' function G, such as a standardised multivariate normal density function (see Lee, 2005, p.193), in which a greater weight would be placed upon schools which were closer to the Academy in terms of their observable characteristics X_{jo}, with the associated weights given by:

$$W_{n_o n_{A(t)}}(j,h) = G_{jh} / \sum_{\ell \varepsilon C} G_{j\ell} \quad \text{where } G_{j\ell} = G((X_{jo} - X_{\ell o}) / \varpi_{n_o}) \qquad (12.21)$$

where $n_{A(t)}$ is again the number of Academies that have been opened by time t, n_o is the number of schools in the comparison group C, and ϖ_{n_o} is the band width used in the matching process.

A special case of (12.21) is matching each Academy j with its **nearest neighbour**. For the first Comparison Group that is defined in terms of average KS2 prior attainment scores, this would be the school h_j that was closest to a given Academy j in its average KS2 prior attainment score in the baseline year. For the further Comparison Group that is based upon a wider set of variables to determine the propensity score matching, it would be the school h_j which is closest in terms of its propensity score $p_h(X)$ identified above. Nearest neighbour matching then involves use of the estimator

$$D_t(R) = \sum_{j \varepsilon A(t)} [(V_{jt} - V_{jo}) - (V_{h_j t} - V_{h_j o})] / n_{A(t)} \tag{12.22}$$

A third form of matching estimator uses local linear matching (Heckman *et al*, 1997, p. 630), using the weights

$$W_{n_o n_{A(t)}}(j,h) = \frac{G_{jh} \sum_{\ell \varepsilon C} G_{j\ell}(X_{\ell s} - X_{js})^2 - [G_{jh}(X_{hs} - X_{js})][\sum_{\ell \varepsilon C} G_{j\ell}(X_{\ell s} - X_{js})]}{\sum_{\ell \varepsilon C} G_{j\ell} \sum_{k \varepsilon C} G_{jk}(X_{ks} - X_{js})^2 - (\sum_{k \varepsilon C} G_{jk}(X_{ks} - X_{js}))^2} \tag{12.23}$$

evaluated at s = 0 or t. A related form of matching that is advocated by Heckman *et al* (1997, p. 630) and Heckman *et al* (1998b, p, 1041), in the context of regression-adjusted difference-in-differences, and conditional difference-in-differences, matching estimators, makes use of local linear matching on the probability of participation p_j in the programme for each school j, with

$$W_{n_o n_{A(t)}}(j,h) = \frac{G_{jh} \sum_{\ell \varepsilon C} G_{j\ell}(p_\ell - p_j)^2 - [G_{jh}(p_h - p_j)][\sum_{\ell \varepsilon C} G_{j\ell}(p_\ell - p_j)]}{\sum_{\ell \varepsilon C} G_{j\ell} \sum_{k \varepsilon C} G_{jk}(p_k - p_j)^2 - (\sum_{k \varepsilon C} G_{jk}(p_k - p_j))^2} \tag{12.24}$$

A fifth, and simpler method, involves the weights:

$$\omega_{n_o n_{A(t)}}(j) = 1/n_{A(t)} \text{ and } W_{n_o n_{A(t)}}(j,h) = 1/n_o \text{ for all } j \varepsilon A(t), h \varepsilon C \tag{12.25}$$

When inserted into the matching estimator (12.20), they result in an evaluation of the **mean effect of treatment on the treated** (Heckman *et al*, 1997, p. 609), corresponding here to an assessment of the average change in the value added that is achieved by the open Academies over the period from the base period up to time t compared to the average change in the value added that is achieved by schools in the comparison group C, as in (12.7) above.

13. EXTENSIONS OF THE EVALUATION

a. Impact on disaggregated measures of educational performance

The analysis in Section 12 of difference-in-differences estimators, and matching procedures, can be applied at a number of different stages of the educational process. These include in particular examination performance at Key Stage 3 (KS3) and Key Stage 4 (KS4), with associated value-added measures from KS2 to KS3, from KS3 to KS4, and from KS2 to KS4. Such performance may be assessed not simply in terms of the overall Average Point Scores of pupils at KS3, or their (capped) total point scores at KS4, but also their disaggregated performance in English, in Mathematics and in Science. In addition, it may include an analysis of post-16 performance for open Academies that have a Sixth Form, including both A-level performance and the acquisition of advanced and intermediate vocational qualifications.

The analysis can be further extended into a comparison of the extent to which the open Academies have succeeded in benefitting all of their pupils, or have tended to benefit some groups of pupils more than others. The relevant groups of pupils may include pupils distinguished by gender, those in ethnic minorities, students from particularly disadvantaged backgrounds, and those with Special Educational Need.

b. Impact on other measures of performance

The analysis in Section 12 of difference-in-differences estimators, and matching procedures, can further extended to a number of other measures of school performance. These other measures include:

i. school attendance, and associated percentage rates of half days missed due to authorised and unauthorised absences;

ii. school levels, and percentage rates, of permanent exclusion of pupils;

iii. the proportion of pupils who stay in education after compulsory school age, and the proportion of pupils who enter further or higher education after Sixth Form studies (where appropriate).

Difference-in-differences estimators for each of the above additional performance measures enable the changes in these performance measures which the open Academies have achieved compared to those of their Predecessor Schools to be themselves compared to those changes which have been achieved over the same period by a matched sample of comparable schools with similar characteristics. All of these changes, moreover, can be analysed against the background of the national trends in these performance measures over the same period of time. In each case, the analysis is dependent, however, upon the availability of reliable data for each relevant year.

c. Impact on pupil intakes

The overall impact of the Academies programme on the educational performance of the schools in the programme can be decomposed into several components. The change, $\Delta\overline{y}_{jt}$, in the mean level, $\overline{y}_{jt}$, of the educational outcome scores y_{ijt} across pupils i in school j at time t, compared to their mean level in the base year before the start of the programme, is itself composed of the following elements:

$$\Delta\overline{y}_{jt} = \Delta V_{jt} + \sum_{k\varepsilon P} \beta_k \Delta\overline{x}_{kjt} + \sum_{h\varepsilon S} \gamma_h \Delta z_{hjt} \qquad (13.1)$$

where $\Delta V_{jt} \equiv V_{jt} - V_{jo}$ for $V_{js} \equiv E_j(V_{ijs})$, $\Delta\overline{x}_{kjt} \equiv \overline{x}_{kjt} - \overline{x}_{kjo}$ for $\overline{x}_{kjs} \equiv E_j(x_{ikjs})$, $\Delta z_{hjt} \equiv z_{hjt} - z_{hjo}$ for s = 0, t, and E_j denotes the mean value of the relevant variable across pupils i in school j. The overall change in the mean level of the educational scores for school j is thus composed of the change in the mean level of the value added V_{ijs} for each pupil i in school j between s = 0 and s = t given by (12.8), plus a weighted sum of the changes which have taken place over the period in the mean values of the pupil characteristics x_{ikjs} within the school and in the school characteristics z_{hjt}. The relevant weights are the coefficients β_k and γ_h of the respective pupil- and school- characteristics in the value-added function in (12.8). For any Academy j that was

open at time t, the corresponding school j at time s = 0 before the Academy opened is its Predecessor School (or the pupil-weighted combination of its Predecessor Schools, if it had more than one Predecessor Schools).

The introduction of the Academies programme may not only affect the mean level of value added which each pupil achieves. By potentially providing more attractive local schools than the Predecessor Schools which they replace, the introduction of the Academies programme may in addition change the nature of the demand for pupil places in the new Academies compared to pattern of demand for pupil places in the corresponding Predecessor Schools. However, in order to assess the impact which the introduction of the Academies programme may have over a period of time since the start of the programme, an allowance must be made for potential changes which may have taken place over the same period of time in the socio-economic characteristics of the areas from which the Academies draw their pupils, due to wider demographic changes.

We will assume that:

$$\bar{x}_{kjt} = a_{kt} + \bar{x}_{kjo} + c_{kjt} + b_{kt}d_{jt} + \mu_{kjt} \quad \text{for } j \varepsilon A(t) \tag{13.2}$$

where a_{kt} is the change in $\bar{x}_{kjs}$ at time s = t compared to time s = 0 that is due to national trends which affect all secondary schools, c_{kjt} is the change in $\bar{x}_{kjs}$ which is due to local demographic changes that affect Academy j, μ_{kjt} is an independently distributed stochastic term, and $d_{jt} = 1$ for those Academies which are open at time t, but $d_{jt} = 0$ otherwise.

In order to identify the impact b_{kt} in (13.2) that is due to the existence of the Academies programme, it is desirable to match the Academies with a Comparison Group that has experienced the same local demographic changes that influence the nature of the pupil intake over the time period. The identification of such a Comparison Group is complicated in practice by the overlapping nature of the local areas from which many schools recruit their pupils, and the lack of any simple geographical template for these overlapping areas. However, one candidate for such a Comparison Group is the cohort of pupils which attended the same Primary Feeder schools as the Predecessor Schools of the Academy in question. Relative weights can be applied to each Primary Feeder to reflect their importance in defining the local

areas from which the Predecessor Schools have recruited their pupils. In order to avoid including in the design of the Comparison Group effects which result from the impact of the Academies Programme, the relative weight on each Primary Feeder can be chosen to be the proportion of the pupil intake into the Academy Predecessor School that came from the Primary Feeder School in the base year of 2001-2 before the start of the Academies Programme.

For each such Primary Feeder school, we will assume that:

$$\bar{x}_{k\ell t} = a_{kt} + \bar{x}_{k\ell o} + c_{k\ell t} + b_{k\ell t} d_{jt} + \mu_{k\ell t} \quad for\ \ell \varepsilon F(j),\ j \varepsilon A(t) \tag{13.3}$$

where *F(j)* is the set of Primary Feeders for the Predecessor Schools of Academy j in the base year. We will assume in the following analysis that the sets $F(j)$ for $j \varepsilon A(t)$ do not overlap, so that their intersections are empty. The case where a school may be a Primary Feeder to the Predecessor Schools of more than one Academy raises additional complications which we will examine in more detail later.

$\bar{x}_{k\ell s}$ in (13.3) is the mean level of pupil characteristic k for pupils in the relevant cohort of pupils from Primary Feeder school ℓ at time s = 0, t. a_{kt} is again a national trend factor which is assumed to apply to all relevant cohorts of pupils for characteristic k between time 0 and time t. We will assume that the local demographic trends $c_{k\ell t}$ in (13.3) and c_{kjt} in (13.2) that affect pupil characteristic k are such that:

$$c_{kjt} = \sum_{\ell \varepsilon F(j)} n_{\ell j} c_{k\ell t} \Big/ \sum_{\ell \varepsilon F(j)} n_{\ell j} \quad for\ j \varepsilon A(t) \tag{13.4}$$

where $n_{\ell j}$ is the number of pupils from the Primary Feeder school ℓ who entered the Predecessor Schools for Academy j in the base year. (13.2) - (13.4) imply that local demographic factors affect the mean level of pupil characteristic k in Academy j to the same extent as they affect a weighted average of the mean levels, $\bar{x}_{k\ell t}$ *for* $\ell \varepsilon F(j)$, of the pupil characteristic k at time t for the cohorts of pupils from the Primary Feeder schools who served Academy j's Predecessor Schools.

We will also assume that the expected values of the stochastic terms in (13.2) and (13.3) are such that:

$$E(\mu_{kjt}) = 0 \;\; \& \;\; E(\mu_{k\ell t}) = 0 \;\; for\ each\ \ell \varepsilon F(j)\ and\ each\ j \varepsilon A(t) \qquad (13.5)$$

From (13.2) – (13.4), we can derive the following **difference-in-differences estimator** of the impact which the introduction of the Academies programme has on the mean level of pupil characteristic k:

$$\hat{b}''_{kt} \equiv (\bar{x}_{kt} - \bar{x}_{ko}) - (\bar{x}'_{kt} - \bar{x}'_{ko}) \;\; where\ \bar{x}'_{ks} \equiv \sum_{j \varepsilon A(s)} [\sum_{\ell \varepsilon F(j)} n_{\ell j} \bar{x}_{k\ell s} / n_t \sum_{\ell \varepsilon F(j)} n_{\ell j}] \;\; for\ s = 0, t \qquad (13.6)$$

and where $\bar{x}_{kt}$ is the mean value of $\bar{x}_{kjt}$ across the set A(t) of n_t open Academies at time t, and $\bar{x}_{ko}$ is the mean value of $\bar{x}_{kjo}$ across the set A(0) of Academy Predecessor Schools (or combination Predecessor Schools, where there is more than one Predecessor School for a given open Academy).

Under condition (13.5), we will have:

$$E(\hat{b}''_{kt}) = b_{kt} - b'_{kt} \qquad where \;\; b'_{kt} \equiv \sum_{j \varepsilon A(t)} (\sum_{\ell \varepsilon F(j)} n_{\ell j} b_{k\ell t} / n_t \sum_{\ell \varepsilon F(j)} n_{\ell j}) \qquad (13.7)$$

If the pupil characteristic k is the prior attainment level at KS2, the relevant cohorts of pupils in the study of the determinants of the change in KS3 examination performance between the base year of 2001-2 and 2004-5 will be those who took KS2 in the Primary Feeder schools in the academic years of 1998-9 and 2001-2 respectively. Since this predates the introduction of the Academies programme, **in this case** we may assume that the overall mean values of the KS2 performance of the relevant cohorts of pupils who left the Primary Feeders will be unaffected by the introduction of the Academies programme. This implies that the corresponding $b_{k\ell t} = 0$ for each $\ell \varepsilon F(j)$ and hence $b'_{kt} = 0$ in (13.3) and (13.7). The difference-in-differences estimator given by (13.6) will then provide a consistent estimate of the overall impact b_{kt} of the Academies programme in (13.2) and (13.7) on the mean value of pupil characteristic k, here the KS2 prior attainment level of the pupils **which the new Academies succeed in attracting**

to the school, after netting out changes in this mean value which are due to national and local demographic trends. This impact will itself not necessarily be zero, since even if the overall mean values of the KS2 performance of the relevant cohorts of pupils who left the Primary Feeders are unaffected by the introduction of the Academies programme, the range of pupils each Academy succeeds in attracting from within the overall distribution of KS2 scores within these cohorts may differ from that of its Predecessor School.

In examining the factors which may influence examination performance at KS3 and KS4 **in future years**, we may be interested in the pupil prior attainment levels at KS2 for the cohort of pupils who have entered the open Academies **since the academic year 2001-2**. The change in the average level of KS2 prior attainment between the cohort of pupils who entered the Academy Predecessor Schools in 2001-2 and the cohort of pupils who entered one of the open Academies in 2004-5 is also of interest in its own right as an indicator of the impact of Academies programme on the nature of the pupil intake which the Academies attract. However, in order to separate out the impact of the Academies programme from the impacts of national and local demographic trends, we again need to formulate the estimator in a difference-in-differences form, such as (13.6), that makes use of a relevant local Comparison Group, such as the cohorts of pupils from the relevant Primary Feeders.

In the case of the cohort of pupils who left the Primary Feeders in 2004-5 (rather than in 2001-2 before the introduction of the Academies programme) to go to either one of the open Academies or another secondary school, it is possible that their overall characteristics were influenced by the introduction of the Academies programme. It is conceivable, for example, that anticipation of a very successful Academy might influence some parents to relocate in its local area rather than elsewhere, and to send their pupils to one of the local Primary Feeders in advance of their intended entry to the new open Academy. In such a case, the overall characteristics of the cohort of pupils who left the Primary Feeders, not just for the new Academy but also for other local secondary schools, might be influenced by the introduction of the Academies programme. In such a case, the coefficients $b_{k\ell t}$ in (13.3) and (13.7), and hence b'_{kt} in (13.7), might be non-zero.

The difference-in-differences estimator (13.6) will then provide a consistent estimator of the impact b_{kt} of the introduction of the Academies programme on characteristic k of the pupil

intake into the open Academies, relative to the more general impact b'_{kt} which its introduction has on the overall characteristics of the cohort of pupils leaving the local Primary Feeders. The difference-in-differences estimator (13.6) may indeed be written in the form:

$$\hat{b}''_{kt} = (\overline{x}_{kt} - \overline{x}'_{kt}) - (\overline{x}_{ko} - \overline{x}'_{ko}) \tag{13.8}$$

involving the change over the period in the mean value of the pupil intake characteristic k into the open Academies **relative to its overall weighted mean value** for the cohort of pupils from the relevant Primary Feeders, compared to this **relative mean value** in the base year for the Academy Predecessor Schools. The second bracketed term in (13.8) would, for instance, be negative if all the Academy Predecessor Schools in the base year had recruited pupils whose average KS2 prior attainment level was below the overall weighted average of the KS2 prior attainment level for the base year's cohort of pupils leaving the relevant Primary Feeders. The first bracketed term in (13.8) would, however, be positive if all the open Academies in year t recruited pupils whose average KS2 prior attainment level was above the overall weighted average of the KS2 prior attainment level for year t's cohort of pupils leaving the relevant Primary Feeders. A combination of a negative second bracketed term and a positive first bracketed term in (13.8) would imply that the overall impact of the Academies programme on the average KS2 pupil intake characteristic would be positive, but again relative to any more general impact b'_{kt} which its introduction has on the overall characteristics of the cohort of pupils who leave the local Primary Feeders for the Academies or other secondary schools.

In order to focus on the main Primary Feeder schools from whom pupils have been recruited, a Primary Feeder school can be identified as a primary school from which at least 5 pupils went on to the secondary school in question in the baseline year of entry. The first Academies opened in September 2002, and their opening may indeed have influenced their recruitment patterns from primary schools in this year, compared to those which previously prevailed for their Predecessor Schools. However, it is also possible that the impending opening of the Academies influenced the recruitment pattern from primary schools into the Predecessor Schools for a year or two in advance of 2002. In order to net out such an influence, a baseline year of entry of 1999 has attractions for comparing the before- and after- effects of the impact of the Academies programme on pupil recruitment from Primary Feeder schools. The baseline then involves the pupil intake from Primary Feeders into the Academy Predecessor Schools in

1999. Any subsequent change in the identity of primary schools from whom significant numbers of pupils are recruited into the open Academies, compared to this baseline, will then influence the magnitude of the impact of the Academies programme which is estimated by the difference-in-differences estimator (13.6).

d. Impact on cohort heterogeneity

A parallel analysis to the above can be applied to the **school-level characteristics** z_{hjt}, which may be influenced by the introduction of the Academies programme, and by other changing **local socio-economic conditions**. In the case of school-level characteristics, such as the KS2 average point score of the cohort, that are essentially the mean values of pupil-level characteristics, the derivation of appropriate difference-in-differences estimators follows in a directly similar way to (13.2) – (13.8) above. In the case of other school-level characteristics, the analysis can be adapted to the particular characteristics. One such case of interest is the standard deviation of KS2 point scores of pupils in the cohort, as a measure of the **variability or heterogeneity** of the prior attainment levels of the school's pupil intake. More generally, the school-level characteristic h may be the standard deviation σ_{hjt} of some corresponding pupil characteristic within the school j at time t. In such a case, we may assume that:

$$\sigma^2_{hjs} = v^2_{hjs} + \varsigma_{hs} d_{js} + \xi_{hjs} \ \textit{where} \ v^2_{hjs} \equiv \Big(\sum_{\ell \varepsilon F(j)} n_{\ell j} v^2_{h\ell js} \Big/ \sum_{\ell \varepsilon F(j)} n_{\ell j} \Big) \ \ \textit{for} \ j \varepsilon A(s) \qquad (13.9)$$

where $v_{h\ell js}$ is the standard deviation across pupils from the Primary Feeder school ℓ of the corresponding pupil characteristic from the mean value of this pupil characteristic for the relevant cohort of pupils in Academy j, ς_{hs} is a constant that reflects the impact of the Academies programme on the school-level characteristic h at time t, and ξ_{hjs} is a stochastic term with a zero expected value, for s = 0, t. We may then achieve a consistent estimate of the programme impact ς_{ht} through use of the difference-in-differences estimator:

$$\hat{\varsigma}_{ht} = (\bar{\sigma}^2_{ht} - \bar{\sigma}^2_{ho}) - (\bar{v}^2_{ht} - \bar{v}^2_{ho}) \quad \textit{with} \ \ E(\hat{\varsigma}_{ht}) = \varsigma_{ht} \qquad (13.10)$$

where $\bar{\sigma}^2_{hs}$ is the mean value of σ^2_{hjs}, and $\bar{v}^2_{hs}$ is the mean value of v^2_{hjs}, across the open Academies for case where the time s = t > 0 , and across their corresponding Predecessor Schools for the case of the base year s = 0.

e. Impact on Primary Feeder schools

One of the 'intermediate' objectives of the Academies programme (PwC, 2003, p. A1) is "to help raise achievement rates of pupils in other local schools, including feeder primary schools, by sharing facilities and expertise within four years of opening". Although one may expect the extent of the influence to be greater with the passage of time, an assessment may be made of the average impact of the Academies programme over the initial evaluation period on the achievement rates of pupils in relevant Primary Feeder schools, through adopting a similar methodology to that described above. We will assume that equation (13.3) again holds for Primary Feeder schools from whom the Academy Predecessor Schools have recruited their pupils.

The derivation of a difference-in-differences estimator in this context will make use of a Comparison Group Ω of Primary Feeder schools which are not affected by the Academies programme. For each Primary Feeder school ℓ in the Comparison Group, we will assume that the mean level of their pupils' characteristic k at time t can be modelled in the form:

$$\bar{x}_{k\ell t} = a_{kt} + \bar{x}_{k\ell o} + c_{k\ell t} + \mu_{k\ell t} \quad if \ \ell \varepsilon \Omega \tag{13.11}$$

$\bar{x}_{k\ell s}$ is again the mean level of pupil characteristic k for pupils in the relevant cohort of pupils from Primary Feeder school ℓ at time s = 0, t, and a_{kt} is a national trend factor which is assumed to apply to all relevant cohorts of pupils for characteristic k between time 0 and time t. Such pupil characteristic may in particular include their Average Point Score at KS2, both overall and in English, Mathematics and Science separately.

We will assume that the local demographic trends $c_{k\ell t}$ in (13.3) and (13.11) that affect pupil characteristic k are such that:

$$(\sum_{j\varepsilon A(t)} n_{\ell j} c_{k\ell t} / n_t \sum_{j\varepsilon A(t)} n_{\ell j}) = (\sum_{\ell \varepsilon \Omega} n_\ell^o c_{k\ell t} / \sum_{\ell \varepsilon \Omega} n_\ell^o) \quad (13.12)$$

where n_ℓ^o is the relevant number of pupils in the Primary Feeder school $\ell \varepsilon \Omega$. (13.12) implies that the pupil-numbers weighted impact of local demographic trends is the same overall for the Comparison Group of Primary Feeder schools as it is for the Academies group of Primary Feeder schools. We may then define the difference-in-differences estimator:

$$\hat{b}_{kt}''' \equiv (\bar{x}_{kt}' - \bar{x}_{ko}') - (\bar{x}_{kt}'' - \bar{x}_{ko}'') \ \ where\ \bar{x}_{ks}'' \equiv \sum_{\ell \varepsilon \Omega} n_\ell^o \bar{x}_{k\ell s} / \sum_{\ell \varepsilon \Omega} n_\ell^o \quad for\ s = 0, t \quad (13.13)$$

and where $\bar{x}_{ks}'$ is defined in equation (13.6). If we also assume that the expected values $E(\mu_{k\ell t})$ of the stochastic terms in (13.3) and (13.11) are zero, the above estimator can be shown to provide an unbiased estimate of the average programme impact b_{kt}''' (given by (13.7)) of the Academies programme on the pupil characteristic k across the relevant Primary Feeder schools, with:

$$E(\hat{b}_{kt}''') = b_{kt}''' \quad (13.14)$$

f. Impact on other secondary schools

The evaluation of the Academies programme may be further extended to the evaluation of other local secondary schools on which they may have an influence. In some areas, the identity of these other secondary schools may be obvious, due to a geographical clustering of schools. However, in other cases, such as large cities, where several Academies are located, the impact of the Academies schools on the performance of other secondary schools in the area may be less clear-cut due to many cross-city linkages between secondary schools and the locations from which pupils originate. A group of schools with whom these changes might be compared is those schools which have had patterns of pupil recruitment from Primary Feeder schools which overlap with the patterns of pupil recruitment from Primary Feeder schools which have prevailed for the Academy Predecessor Schools. A Primary Feeder school is again identified in this analysis as a primary school from which at least 5 pupils went on to the secondary school

in question in the baseline year of entry, here 1999, after taking Key Stage 2 in the primary school. A school is then defined as an Overlapping Intake School (OIS) to an Academy Predecessor School if both it and the Academy Predecessor School recruited at least 10 pupils from amongst their common Primary Feeder schools in the baseline year of entry, here 1999 for the cohort of pupils who went on to take GCSE in 2004. A search of the PLASC database enables the identity of the Primary Feeder schools that send, or have in the recent past sent, 10 or more pupils to a given Academy or its predecessor school to be identified. Similarly it enables all the other secondary schools which are or have recently been the recipients of 10 or more pupils from the same feeder primary schools to be identified. These secondary schools are then competing with the Academy for the pupils from Primary Feeder schools which have supplied in the recent past both this OIS group of schools and the Predecessor Schools of the new Academy. Such a competitive spur may indeed provide one of the main incentives for the OIS group of schools to improve their educational effectiveness.

One main test of whether the introduction of the Academies programme has had a significant effect upon the **educational effectiveness** of other secondary schools with overlapping sources of pupil intake is an assessment of the direction and magnitude of the impact which this programme has had on the **educational value added** of the OIS schools, **after adjusting for the changes in the pattern of pupil recruitment** which the Academies programme may have influenced. The impact of the introduction of the Academies programme on the educational attainment of pupils in the OIS schools may be modelled in a parallel way to equation (12.1) above, but now using the equation

$$y_{ijt} = \alpha_t + \sum_{k\varepsilon P} \beta_{kt} x_{kijt} + \sum_{h\varepsilon S} \gamma_{ht} z_{hjt} + \zeta_t m_{jt} + \theta_{jt} + \varepsilon_{ijt} \quad for \quad j \notin A(t) \qquad (13.15)$$

where m_{jt} is the number of Academies at time t for which school j was an OIS school. Equation (13.15) allows for the possibility of overlapping sets of OIS schools for different Academies, as may occur if there is a geographical concentration of more than one Academy in a region, as well as for cases where $m_{jt} = 0$. In a similar way to Bettinger (2005), equation (13.15) assumes that the stimulus given by Academies is proportional to the number of

Academies for which a given school has been an OIS school. Other non-linear formulations of their influence in (13.15) are indeed possible.

The overall impact which the Academies programme has on educational attainment may be decomposed in a similar way to equation (13.1) above, into those changes which are associated with changes in the characteristics of the pupil intake, and in school-level inputs, and those changes which are due to improvements in the value added achieved by the OIS group of schools as a result of the stimulus associated with the introduction of the Academies programme. We may then define the regression-adjusted conditional difference-in-differences estimator

$$\hat{\zeta}_t = [(V_t^O - V_o^O) - (V_t^C - V_o^C)]\ [n_{\Phi(t)} / \sum_{j \varepsilon \Phi(t)} m_{jt}\ ] \qquad (13.16)$$

where each V_s term is defined as in (12.7), $\Phi(t)$ is the set of OIS schools for the open Academies at time t, and $n_{\Phi(t)}$ is the number of such OIS schools at time t. The estimator (13.6) compares the improvement in the mean value added for the OIS group of schools, denoted by superscript *0*, with the improvement in the mean value added for a comparison group C of schools that are neither OIS schools nor Academies, per average number of Academies for which the OIS schools are OIS schools. Under similar conditions to those discussed earlier, the estimator $\hat{\zeta}_t$ will provide a consistent estimate of the impact parameter ζ_t in equation (13.15).

The impact which the introduction of the Academies programme has on patterns of pupil recruitment can be assessed in a similar way to that discussed in Section 13c. above. Because the changes in (13.6) and (13.8) involve changes over the period in the mean value of the pupil intake characteristics into the open Academies relative to their overall weighted mean values for the cohort of pupils from the corresponding Primary Feeders, they also reflect complementary changes associated with the introduction of the Academies programme in the pattern of pupil recruitment for overlapping intake schools, who recruit from the same set of overlapping Primary Feeders as the open Academies.

A high geographical concentration of Academies can also give rise to the additional complication that one of the OIS schools is itself an Academy. Equation (13.15) can be extended to include this possibility through the formulation:

$$y_{ijt} = \alpha_t + \sum_{k\varepsilon P} \beta_{kt} x_{kijt} + \sum_{h\varepsilon S} \gamma_{ht} z_{hjt} + d_{jt}\delta_t + \zeta_t m_{jt} + \theta_{jt} + \varepsilon_{ijt} \qquad (13.17)$$

where again $d_{jt} = 1$ denotes that school j is an Academy at time t, and $d_{jt} = 0$ denotes that it is not, and where m_{jt} now refers to the number of other schools in the OIS group for school j that are Academies. Equation (13.17) itself assumes an additive effect of influence of Academy status for the school itself and the influence on it of any other Academies whose Predecessor Schools have recruited from overlapping Primary Feeders to those of the first Academy. Whilst more complicated interaction terms between these different influences might also be included in (13.17), a low degree of confidence is likely to be attached to their estimation whenever the number of Academies that are also overlapping intake schools to other Academies is small.

Further complications can arise because of the existence of other intervention programmes, such as the Excellence in Cities programme (Machin, McNally and Meghir, 2004), that are similarly aimed at improving the performance of disadvantaged schools. Where some schools are members of more than one such programme, such complications are further compounded by the extent of their interactive effects, which may be best estimated by regarding such combined membership of more than one programme as an additional form of treatment in the list of possible 'multiple treatments'. As noted by Hsu (1996), a range of pairwise comparisons may be made between the relative impact of different pairs of treatment. The simplest in the present context is the pairwise comparison between the relative impact of the Academies programme and that of a control group in similar circumstances but which are not members of another relevant programme. This would involve excluding schools from the Comparison Groups discussed above that have been members of other relevant programmes, such as Excellence in Cities. However, if the impact of the Academies programme on other secondary schools is to be isolated from the influence of the other programmes on the other secondary schools, it would also involve excluding schools that have been members of other relevant programmes from the set of OIS schools that are considered in (13.15) – (13.17), and reducing the sample size involved. Only the average impact of the Academies programme on the OIS schools which were not in the other programmes may then be identified, though more complex

methods of propensity score matching (see Lee, 2005, pp. 176-7) might identify wider population effects. Imbens (2000) and Lechner (2001) show that average treatment effects (ATEs) can be identified under a multiple-programme version of the **conditional independence assumption (CIA)**, that for any vector of individual characteristics X the outcomes of all potential treatments are independent of how individuals are selected into the different programmes. Using this version of the CIA, Lechner (2001, 2002) provides a matching estimator to estimate the causal treatment effect of each programme based upon the estimation of individual pairwise conditional programme participation probabilities and the pairwise comparison of conditional programme outcomes.

If the list of other relevant programmes is considered to be a long one, through inclusion of programmes such as Education Action Zones (EAZs) that may have had an impact on some secondary schools in similar circumstances, the task of finding a large control group of schools which have comparable characteristics (or similar propensity scores for being chosen as Academies) to the Academies themselves, but which have not been members of these other programmes, becomes a more difficult one. This becomes even more the case if some schools have been members of more than one programme, since each distinct combination of different programmes might be regarded as a different treatment, unless their effects are taken as purely additive.

Against the background of a large number of policy initiatives which may have had an impact on school performance over several years, an alternative approach is to seek to estimate the average impact of the Academies programme relative to a comparison group made up of a representative sample of schools with similar characteristics, or programme participation propensity scores, to the Academies, who have been subject to a representative range of such other programme initiatives. The objective of a new policy initiative, such as the Academies programme, may indeed be viewed as to achieve better results than what in general has preceded it for schools with similar characteristics, or similar chances of being selected for the programme. The use of such a representative comparison group may then provide an appropriate means of evaluating whether such an objective has been achieved.

14. CONCLUSION

Value-added analysis provides an important methodology for assessing the contribution which schools make to the educational attainment of their pupils. It can moreover be productively linked to programme evaluation techniques in contexts where randomised control trials are not feasible, particularly through the use of regression-adjusted conditional difference-in-differences estimators. By adjusting measures of pupil attainment, such as examination results, for pupil prior attainment and other relevant pupil- and school-level variables, value-added analysis not only isolates more closely the contribution which the individual school makes to the pupil's educational progress, but at the same time corrects for many of the factors which would otherwise bias estimates of the impact which participation in an educational initiative, such as the Academies programme, has on those schools in the programme.

Value-added analysis using multilevel modelling takes into account the hierarchical structure of educational data in order to produce more efficient estimates of value added by individual schools than those produced by OLS multivariate regression analysis of pupil-level data. The use of pupil-level data itself has considerable statistical advantages over reliance upon purely school-level mean data to assess changes in school effectiveness. Through its deployment of multilevel modelling and adjustment for many relevant pupil- and school-level variables, the DfES's own Contextual Value Added estimates represent a significant advance compared to earlier non-parametric value-added estimates.

By combining value-added analysis with relevant difference-in-differences estimators, the impact of the Academies programme can be explored in several important directions that are discussed above. Given the important contribution which value-added analysis can make both to programme evaluation and to the assessment of school effectiveness, there is also a need for further research more widely into the impact which factors such as endogeneity bias, measurement error, choice of functional form and parsimony in the selection of explanatory variables, can make to value-added estimates and their robustness, and into the relative merits of different estimation techniques in the face of these additional considerations. While any conclusions based upon existing value-added models will be contingent upon the assumptions implicit in them, such further research can advance our existing state of knowledge of the effect of possible departures from these underlying assumptions.

REFERENCES

Aakkvid, A., Heckman, J. and Vytlacil, E. (2005). Estimating treatment effects for discrete outcomes when responses to treatment vary: an application to Norwegian vocational rehabilitation programs. *Journal of Econometrics*, vol. 125, pp. 15 - 51.

Aitkin, M. and Longford, N.T. (1986). Statistical modelling issues in school effectiveness studies. *Journal of the Royal Statistical Society*, Series A, vol. 149, pp. 1 – 43.

Aitkin, M. and Zuzovsky, R. (1992). New paradigm for the analysis of hierarchically structured data in school effectiveness studies. Paper presented to AERA annual meeting, San Francisco.

Angrist, J. (2004). Treatment effect heterogeneity in theory and practice. *Economic Journal*, vol. 114, pp. C52-C83.

Ashenfelter, O. (1978). Estimating the effect of training programs on earnings. *Review of Economics and Statistics*, vol. 60, pp. 47 – 57.

Battese, G. and Coelli, T. (1995). A model for technical inefficiency effects in a stochastic frontier production function for panel data. *Empirical Economics*, vol. 20, pp. 325-332.

Becker, G. (1993). *Human Capital* (3rd edn). University of Chicago Press: Chicago.

Belfield, C. (2000). *Economic Principles for Education*. Edward Elgar: Cheltenham.

Bell, B., Blundell, R. and Van Reenan (1999). Getting the unemployed back to work: an evaluation of the New Deal proposals. *International Tax and Public Finance*, vol. 6, pp. 339 – 360.

Bettinger, E. (2005). The effect of charter schools on charter students and public schools. *Economics of Education Review*, vol. 24, pp. 133-147.

Blundell, R. and Costa Dias, M. (2000). Evaluation methods for non-experimental data. *Fiscal Studies*, vol. 21, pp. 427 – 468.

Borich, G. (1996). *Effective Teaching Methods*, 3rd edn. Macmillan: New York.

Bound, J., Brown, C. and Mathiowetz, N. (2001). Measurement error in survey data. In Heckman, J. and Leamer, E. (eds.). *Handbook of Econometrics*, vol. 5, pp. 3705-3836.

Box, G. and Cox, D. (1964). An analysis of transformations. *Journal of the Royal Statistical Society*, Series B, vol. 26, pp. 211 - 252.

Bradley, S. and Taylor, J. (1998). The effect of school size on exam performance in secondary schools. *Oxford Bulletin of Economics and Statistics*, vol. 60, pp. 291–324.

Brandsma, H. and Knuver, J. (1989). Effects of school and classroom characteristics on pupil progress in language and arithmetic. *International Journal of Educational Research*, vol. 13, pp. 777 – 788.

Brimer, M., Madaus, G., Chapman, B., Kellaghan, T. and Wood, R. (1977). *Sources of Difference in School Attainment*. Carnegie Corporation: New York.

Browne, W., Goldstein, H., Woodhouse, G. and Yang, M. (2001). An MCMC algorithm for adjusting for errors in variables in random slopes multilevel models. *Multilevel Modelling Newsletter*, vol. 13, pp. 4 – 10.

Bryson, A., Dorsett, R. and Purdon, S. (2002). *The Use of Propensity Score Matching in the Evaluation of Active Labour Market Policies.* Working Paper No. 4, Department of Work and Pensions: London.

Card, D. and Krueger, A. (1992a). Does school quality matter? Returns to education and the characteristics of public schools in the United States. *Journal of Political Economy*, vol. 100, pp. 1 - 40.

Card, D. and Krueger, A. (1992b). School quality and black-white relative earnings: a direct assessment. *Quarterly Journal of Economics*, vol. 107, pp. 151-200.

Chambers, R. G. (1988). *Applied Production Analysis*. Cambridge University Press: Cambridge.

Chambers, T. , Smith, H., Blackburn, B., Silverman, B., Schroeder, B., Reitman, D. and Amroz,. A. (1981). A method for assessing the quality of a randomized control trial. *Controlled Clinical Trials: design, methods and analysis*, vol. 2, pp. 31-49.

Coleman, J.S. (1975). Methods and results in the IEA studies of the effects of school on learning. *Review of Educational Research*, vol. 45, pp. 335 – 386.

Coleman, J.S. and others (1966), *Equality of Educational Opportunity*. United States Department of Health, Education, and Welfare: Washington, DC.

Coleman, J.S., Hoffer, T. and Kilmore, S. (1982). *High School Achievement*. Basic Books: New York.

Cooper, W. W. , Seiford, L. and Tone, K. (2000). *Data Envelopment Analysis*. Kluwer: Boston.

Cullingford, C. and Daniels, D. (1999). Effects of Ofsted inspections on school performance. In Cullingford, C. (ed.). *An Inspector Calls: OFSTED and its Effect on School Standards*. Kogan Page: London. pp. 59 - 69.

Daly, P. (1986). *School Effectiveness and Pupils' Examination Performance in Northern Ireland*. Queen's University: Belfast.

Davidson, R. and MacKinnon, J. (1981). Several tests for model specification in the presence of alternative hypotheses. *Econometrica*, vol. 49, pp. 781-793.

Davidson, R. and MacKinnon, J. (1993). *Estimation and Inference in Econometrics*. Oxford University Press: Oxford.

Davidson, R. and MacKinnon, J. (2004). *Econometric Theory and Methods*. Oxford University Press: Oxford.

Dearden, L., McIntosh, S., Myck, M. and Vignoles, A. (2000). *The Returns to Academic, Vocational and Basic Skills in Britain*, Skills Task Force Research Paper SKT27. Centre for Economic Performance: London.

Dearing, R. (1993). *The National Curriculum and Its Assessment: Interim Report*. School Curriculum and Assessment Authority: London.

Dehejia, R. (2005). Practical propensity score matching: a reply to Smith and Todd. *Journal of Econometrics*, vol. 125, pp. 355-364.

Dehejia, R. and Wahba, S. (2002). Propensity score-matching methods for non-experiemental causal studies. *Review of Economics and Statistics*, vol. 84, pp. 151-161.

Department for Education and Skills (2001). *Statistics of Education: Pupil Progress in Schools in England: 2000.* National Statistics Bulletin: London.

Department for Education and Skills (2002). *Statistics of Education: Pupil Progress in Secondary Schools by School Type in England: 2001.* National Statistics Bulletin: London.

Department for Education and Skills (2003a). *Secondary School Performance Tables 2002.* DfES: London (http://www.dfes.gov.uk/performancetables/schools_02/sec3b.shtml).

Department for Education and Skills (2003b). *Statistics of Education: Pupil Progress by Pupil Characteristics: 2002*. National Statistics Bulletin: London.

Department for Education and Skills (2003c). What are Academies? http://www.standards.dfes.gov.uk/academies/what_are_academies/version-1.

Department for Education and Skills (2005). *Contextual Value Added Information*. http://www.standards/dfes.gov.uk/performance/1316337/CVAinPAT2005/?version=1.

Department for Education and Skills (2006). *Consistent Financial Reporting*. http://www.dfes.gov.uk/valueformoney/index.cfm?action=CFR.Default

Dolton, P. and Vignoles, A. (2002). The return on post-compulsory school mathematics study. *Economica*, vol. 69, pp. 113 – 141.

Farrell, M. (1957). The measurement of productive efficiency. *Journal of the Royal Statistical Society*, Series A, vol. 120, pp. 253-282.

Feinstein, L. and Symons, J. (1999), Attainment in secondary school, *Oxford Economic Papers*, vol. 51, 300-321.

Ferguson, R. and Ladd, H. (1996). How and why money matters: an analysis of Alabama schools. In Ladd, H. (ed.), *Holding Schools Accountable: Performance-Based Reform in Education*. Brookings Institution: Washington, DC, pp. 265-298.

Fernandez, C., Koop, G. and Steel, M. (2005). Alternative efficiency measures for multiple-output production. *Journal of Econometrics*, vol. 126, pp. 411-444.

Fitz-Gibbon, C.T. (1991). Multilevel modelling in an indicator system. In Raudenbush, S. and Willms, J.D. (eds.). *Schools, Classrooms, and Pupils: International Studies of Schooling from a Multilevel Perspective*. Academic Press: London, pp. 37 -51.

Fitz-Gibbon, C.T. (1995). *The Value Added National Project: Issues to Be Considered in the Design of a National Value Added System*. SCAA: London.

Fitz-Gibbon, C.T. (1996). *Monitoring Education: Indicators, Quality and Effectiveness*. Cassell: London.

Fitz-Gibbon, C.T. (1997). *The Value Added National Project Final Report – Feasibility Studies for a National System of Value-Added Indicators*. SCAA: London.

Fitz-Gibbon, C.T. and Morris, L. (1987). *How to Design a Program Evaluation*. Sage: London.

Fuller, W. (1987). *Measurement Error Models*. John Wiley: New York.

Gamoran, A. (1991). Schooling and achievement: additive versus interactive models. In Raudenbush, S. and Willms, J.D. (eds.). *Schools, Classrooms, and Pupils: International Studies of Schooling from a Multilevel Perspective*. Academic Press: London, pp. 37 -51.

Goldhaber, D. and Brewer, D. (1998). Why don't schools and teachers seem to matter? assessing the impact of unobservables on educational productivity. *Journal of Human Resources*, vol. 32, pp. 505-523.

Goldhaber, D., Brewer, D. and Anderson, D. (1999). A three-way error components analysis of educational productivity. *Education Economics*, vol. 7, pp. 199 – 208.

Goldstein, H. (1987). *Multilevel Models in Educational and Social Research*. Griffin: London.

Goldstein, H. (1991). Better ways to compare schools? *Journal of Educational Statistics*, vol. 16, pp. 89 – 91.

Goldstein, H. (1995). *Multilevel Statistical Models*, 2nd edn. Arnold: London.

Goldstein, H. (2001). Using pupil performance data for judging schools and teachers: scope and limitations. *British Educational Research Journal*, vol. 27, pp. 433 – 442.

Goldstein, H. and Sammons, P. (1997). The influence of secondary and junior schools on sixteen year examination performance. *School Effectiveness and School Improvement*, vol. 8, pp. 219 – 230.

Gray, J. (1989). Multilevel models: issues and problems emerging from their recent application in British studies of school effectiveness. In Bock, R.D. (ed.), *Multilevel Analysis of Educational Data*, Academic Press: London, pp. 127- 145.

Gray, J., Goldstein, H. and Jesson, D. (1996). Changes and improvements in schools' effectiveness: trends over five years. *Research Papers in Education*, vol. 11, pp. 35-51.

Gray, J., Goldstein, H. and Thomas, S. (2001). Predicting the future: the role of past performance in determining trends in institutional effectiveness at A level. *British Educational Research Journal*, vol. 27, pp. 391-405.

Gray, J., Jesson, D., Goldstein, H., Hedger, K. and Rasbash, J. (1995). A multi-level analysis of school improvement: changes in schools' performance over time, *School Effectiveness and School Improvement*, vol. 6, pp. 97 – 114.

Gray, J., Jesson, D., and Sime, N. (1995). Estimating differences in the examination performances in secondary schools in six LEAs: a multi-level approach to school effectiveness. In Gray and Wilcox (1995), pp. 105-129.

Gray, J. and Wilcox, B. (1995). *'Good School, Bad School'*. Open University Press: Buckingham.

Greene, W. (2004). Distinguishing between heterogeneity and inefficiency: stochastic frontier analysis of the World Health Organisation's panel data on national health care systems. *Health Economics*, vol. 13, pp. 959-980.

Greene, W. (2005). Reconsidering heterogeneity in panel data estimators of the stochastic frontier model. *Journal of Econometrics*, vol. 126, pp. 269 – 303.

Gujarati, D.N. (1995). *Basic Econometrics*, 3rd edn. McGraw-Hill: New York.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, vol. 66, pp. 315-331.

Hanushek, E., Rivkin, S. and Taylor, L. (1996). Aggregation and the estimated effects of school resources. *Review of Economics and Statistics*, vol. 78, pp. 611-627.

Hargreaves, D. (2001). A capital theory of school effectiveness and improvement. *British Educational Research Journal*, vol. 27, pp. 487 – 503.

Heath, A. and Clifford, P. (1981). The measurement and explanation of school differences. *Oxford Review of Education*, vol. 7, pp. 33 – 40.

Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica*, vol. 47, pp. 153 – 161.

Heckman, J., Ichimura, H. and Todd, P. (1997). Matching as an econometric estimator: evidence from evaluating a job training programme. *Review of Economic Studies*, vol. 64, pp. 605 – 654.

Heckman, J., Ichimura, H. and Todd, P. (1998a). Matching as an econometric estimator. *Review of Economic Studies*, vol. 64, pp. 261 – 294.

Heckman, J., Ichimura, H., Smith, J. and Todd, P. (1998b). Characterising selection bias using experimental data. *Econometrica*, vol. 66, pp. 1017 – 1098.

Heckman, J. and Smith, J. (1999). The pre-programme earnings dip and the determinants of participation in a social programme: implications for simple programme evaluation strategies. *Economic Journal*, vol. 109, pp. 313 – 348.

Hill, P. and Rowe, K. (1996). Multilevel modelling in school effectiveness research. *School Effectiveness and School Improvement*, vol. 7, pp. 1- 34.

Hopkins, D. and Reynolds, D. (2001). The past, present and future of school improvement: towards the third age. *British Educational Research Journal*, vol. 27, pp. 459 – 475.

Hsu, J. (1996). *Multiple Comparisons: Theory and Methods*. Chapman Hall: London.

Hutchison, D., Morrison, J. and Felgate, R. (2003). Bootstrapping the effects of measurement errors. *Multilevel Modelling Newsletter*, vol. 15, pp. 2 – 10.

Imbens, G. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika*, vol. 87, pp. 706 – 710.

Jencks, C. *et al* (1972). *Inequality: A Reasessment of the Effect of Family and Schooling in America*. Basic Books: New York.

Jesson, D. (2001) *Educational Outcomes and Value Added by Specialist Schools for the Year 2000*. Technology Colleges Trust: London.

Jesson, D. (2002). *Value Added and the Benefits of Specialism*. Technology Colleges Trust: London.

Jesson, D. (2003). *Educational Outcomes and Value Added by Specialist Schools – 2002 Analysis*. Specialist Schools Trust, London.

Jesson, D. (2004). *Educational Outcomes and Value Added by Specialist Schools – 2003*. Specialist Schools Trust, London.

Jesson, D. and Crossley, D. (2005). *Educational Outcomes and Value Added by Specialist Schools – 2004*. Specialist Schools and Academies Trust, London.

Jesson, D. and Crossley, D. (2006). *Educational Outcomes and Value Added by Specialist Schools – 2005*. Specialist Schools and Academies Trust, London.

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

Johnston, J. (1987). *Econometric Methods*, 3rd edn. McGraw-Hill: London.

Jondrow, J., Lovell, K. , Materov, I. S. and Schmidt, P. (1982). On the estimation of technical efficiency in the stochastic frontier production function model. *Journal of Econometrics*, vol. 19, pp. 233 – 238.

Kay, J. (1993). *Foundations of Corporate Success*. Oxford University Press: Oxford.

Kenny, D. (1975). Quasi-experimental approach to assessing treatment effects in the nonequivalent control group design. *Psychological Bulletin*, vol. 83, pp. 345 – 362.

Kreft, I. (1996). Are multilevel techniques necessary? An overview, including simulation studies. http://www.calstatela.edu/faculty/ikreft/quarterly/quarterly.html

Kreft, I., Leeuw, J. de and Leeden, R van der (1994). Review of five multilevel analysis programs. *The American Statistician*, vol. 48, pp. 324 – 335.

Kumbhakar, S. and Lovell, K. (2000). *Stochastic Frontier Analysis*. Cambridge University Press: Cambridge.

Lechner, M. (2001). Identification and estimation of causal effects of multiple treatments under the conditional independence assumption. In Lechner, M. and Pfeiffer, F. (eds.), *Econometric Evaluation of Active Labor Market Policies in Europe*. Physica/Springer: Heidelberg, pp. 43-58.

Lechner, M. (2002). Program heterogeneity and propensity score matching: an application to the evaluation of active labor market policies. *Review of Economics and Statistics*, vol. 84, pp. 205-220.

Lee, M.-J. (2005). *Micro-Econometrics for Policy, Program and Treatment Effects*. Oxford University Press: Oxford.

Leeuw, J. de and Kreft, I. (1995). Questioning multilevel models. *Journal of Educational and Behavioral Statistics*, vol. 20, pp. 171 - 189.

Levacic, R. and Woods, P. (2002a). Raising school performance in the league tables (part 1): disentangling the effects of social disadvantage. *British Educational Research Journal*, vol. 28, pp. 207 – 226.

Levacic, R. and Woods, P. (2002b). Raising school performance in the league tables (part 2): barriers to responsiveness in three disadvantaged schools. *British Educational Research Journal*, vol. 28, pp. 227 – 247.

Machin, S., McNally, S. and Meghir, C. (2004). Improving pupil performance in English secondary schools: Excellence in Cities. *Journal of the European Economic Association*, vol. 2, pp. 396-405.

Marks, J., Cox, C. and Pomian-Srzednicki, M. (1983). *Standards in English Schools*. National Council for Educational Standards: London.

Mayston, D.J. (2000). The demand for education and the production of local public goods. *Discussion Papers in Economics*, 2000/50. University of York: York.

Mayston, D.J. (2002). *Tackling the Endogeneity Problem when Estimating the Relationship Between School Spending and Pupil Outcomes*. DfES Research Report RR328. Department for Education and Skills: London.

Mayston, D.J. (2003a). Value for money, educational resourcing and pupil attainment. In *Public Services Productivity*, Proceedings of Seminar on Productivity in Public Services held at HM Treasury, June 2002. HM Treasury: London, pp. 27 – 31.

Mayston, D.J. (2003b). Measuring and managing educational performance. *Journal of the Operational Research Society*, vol. 54, pp. 679-691.

Mayston, D. J. (2007a). Competition and resource effectiveness in education. *Manchester School*, vol. 75, no. 1, pp. 47 – 64.

Mayston, D. J. (2007b). Adding more value to educational value added. *Discussion Papers in Economics*, forthcoming. University of York: York.

Mayston, D.J. and Jesson, D. (1988). Developing models of educational accountability. *Oxford Review of Education*, vol. 14, pp. 321-340.

Mayston, D.J. and Jesson, D. (1999). *Linking Educational Resourcing with Enhanced Educational Outcomes*. DfEE Research Report 179. Department for Education and Employment: London.

McCullagh, P. (1989). What can go wrong with iteratively re-weighted least squares? In Bock, R.D. (ed.), *Multilevel Analysis of Educational Data*, Academic Press: London, pp. 147 - 156.

Montgomery, J., Byerly, M., Carmody, T. , Li, B., Miller, D., Varghese, F. and Holland, R. (2004). An analysis of the effect of funding sources in randomized control trials of second generation antisychotics for the treatment of schizophrenia. *Controlled Clinical Trials: design, methods and analysis*, vol. 25, pp. 598-612.

Montmarquette, C. and Mahseredjian, S. (1985). Functional forms and educational production functions. *Economic Letters*, vol. 19, 291-294.

Montmarquette, C. and Mahseredjian, S. (1989). Does school matter for educational achievement? A two-way nested error components analysis. *Journal of Applied Econometrics*, vol. 4, 181-193.

Mood, A. and Graybill, F. (1963). *Introduction to the Theory of Statistics* (2nd edn.). McGraw-Hill: New York.

Morris, C. (1995). Hierarchical models for educational data: an overview. *Journal of Educational and Behavioral Statistics*, vol. 20, pp. 190 - 200.

Mortimore, P., Sammons, P., Stoll, L., Lewis, L. and Ecob, R. (1988). *School Matters: The Junior Years*. Open Books: Wells.

Newton, P. (2005). The public understanding of measurement inaccuracy. *British Educational Research Journal*, vol. 31(4), pp. 419-442.

Nutall, D., Goldstein, H., Prosser, R. and Rasbash, J. (1989). Differential school effectiveness. *International Journal of Educational Research* vol. 13(7), pp. 769-776.

Office for Standards in Education (2000). *Improving City Schools*, HMI 222. OFSTED: London.

Office for Standards in Education (2003). *Excellence in Cities and Educational Action Zones: management and impact,* HMI 1399. OFSTED: London.

Plewis, I. (1991). Using multilevel models to link educational progress with curriculum coverage. In Raudenbush, S. and Willms, J.D. (eds.). *Schools, Classrooms, and Pupils: International Studies of Schooling from a Multilevel Perspective*. Academic Press: London, pp. 53 - 65.

Postlethwaite, T. (1975). The surveys of the IEA. In Purvis, A. and Levine, D. (eds.). *Educational Policy and International Assessment*. McCutchen: Berkeley.

Preece, P. (1989). Pitfalls in research on school and teacher effectiveness. *Research Papers in Education*, vol. 4, pp. 47 – 69.

PricewaterhouseCoopers (2003), *Department for Education and Skills - Academies Evaluation Annual Report: Annexes*. PricewaterhouseCoopers: London.

Pugh, G. and Mangan, J. (2003). What's in a trend? A Comment of Gray, Goldstein and Thomas (2001). *British Educational Research Journal*, vol. 29, pp. 77-82.

Puhani, p. (2000). The Heckman correction for sample selection and its critique - a short survey, *Journal of Economic Literature*, vol. 14, pp. 53 – 68.

Raffe, D. (1991). Assessing the impact of a decentralised initiative: the British Technical and Vocational Education Initiative. In Raudenbush, S. and Willms, J.D. (eds.). *Schools, Classrooms, and Pupils: International Studies of Schooling from a Multilevel Perspective*. Academic Press: London, pp. 149 - 166.

Raudenbush, R. (1989a). 'Centering' predictors in multilevel analysis: choices and consequences. *Multilevel Modelling Newsletter*, vol. 1 (2), pp. 10 – 12.

Raudenbush, R. (1989b). The analysis of longitudinal, multilevel data. *International Journal of Educational Research* vol. 13(7), pp. 721-740.

Raudenbush, R. and Bryk, A. (1989). Quantitative models for estimating teacher and school effectiveness. In Bock, R.D. (ed.), *Multilevel Analysis of Educational Data*. Academic Press: London, pp. 205-232.

Reynolds, D., Hopkins, D. Potter, D., and Chapman, C. (2001). *School Improvement for Schools Facing Challenging Circumstances*. DfEE: London.

Reynolds, D. and Teddlie, C. (2000). The processes of school effectiveness. In Teddlie, C. and Reynolds, D. (eds.). *The International Handbook of School Effectiveness Research*. Falmer Press: London, pp. 134 – 159.

Rice, N., Jones, A. and Goldstein, H. (1999). Multilevel models where the random effects are correlated with the fixed predictors: a conditional iterative generalised least squares estimator (CIGLS). Centre for Health Economics: York.

Rosenbaum, P. and Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, vol. 70, pp. 41 – 55.

Rutter, M. (1983). School effects on pupil progress: research findings and policy implications. *Child Development*, vol. 54, pp. 1-29.

Rutter, M., Maugham, B., Mortimore, P. and Outon, J. (1979). *Fifteen Thousand Hours: Secondary Schools and Their Effects on Children*. Open Books: London.

Sammons, P. (1995). Gender, ethnic and socio-economic differences in attainment and progress: a longitudinal analysis of student achievement over 9 years. *British Educational Research Journal*, vol. 21, pp. 465 – 485.

Sammons, P. (1999). *School Effectiveness: Coming of Age in the Twenty-First Century*. Swets & Zeitlinger: Lisse.

Sammons, P., Hillman, J. and Mortimore, P. (1995). *Key Characteristics of Effectiveness Schools*. OFSTED: London.

Sammons, P., Nuttall D. and Cuttance, P. (1993), Differential school effectiveness: results from a reanalysis of the Inner London Education Authority's junior school project data, *British Educational Research Journal*, vol. 19, pp. 381 – 405.

Sammons, P., Mortimore, P. and Thomas, S. (1996). Do schools perform consistently across outcomes and areas? In Gray, J., Reynolds, D., Fitz-Gibbon, C. and Jesson, D. (eds.) *Merging Traditions: The Future of Research on School Effectiveness and School Improvement*. Cassell: London.

Sanders, W. and Horn, S. (1994). The Tennessee value-added assessment system (TVAAS): mixed-model methodology in educational assessment. *Journal of Personnel Evaluation in Education*, vol. 8, pp. 299-311.

Sanders, W. and Horn, S. (1998). Research findings from the Tennessee value-added assessment system (TVAAS) database: implications for educational evaluation and research. *Journal of Personnel Evaluation in Education*, vol. 12, pp. 247-256.

Shaw, I., Newton, D., Aitkin, A. and Darnell, R. (2003). Do OFSTED inspections of secondary schools make a difference to GCSE results? *British Journal of Educational Research*, vol. 29, pp. 63 – 75.

Shaycroft, M. (1967). *The High School Years: Growth in Cognitive Skills*. American Institute for Research Pittsburgh.

Sickle, R. (2005). Panel estimators and the identification of firm-specific efficiency levels in parametric, semiparametric and nonparametric settings. *Journal of Econometrics*, vol. 126, pp. 305-334.

Simar, L. and Wilson, P. (2000). A general methodology for bootstrapping in nonparametric frontier models. *Journal of Applied Statistics*, vol. 27, pp. 779-802.

Smith, D.J. and Tomlinson, S. (1989). *The School Effect*. Policy Studies Institute: London.

Smith, J. and Todd, P. (2005a). Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of Econometrics*, vol. 125, pp. 305-353.

Smith, J. and Todd, P. (2005b). Rejoinder. *Journal of Econometrics*, vol. 125, pp. 365-375.

Smith, M. (1972). The basic findings reconsidered. In Mosteller, F. and Moynihan, P. (eds.) *On Equality of Educational Opportunity*. Random House: New York.

Spencer, N.H. and Fielding, A. (2002). A comparison of modelling strategies for value-added analyses of educational data, *Computational Statistics*.

Stoll, L. and Myers, K. (ed.) (1998). *No Quick Fixes: Perspectives on Schools in Difficulty*. Falmer Press: London.

Stoney, S., West, A., Kendall, L. and Morris, M. (2002). *Evaluation of Excellence in Cities: Overview of Interim Findings*. http://www.nfer.ac.uk

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

Teddlie, C., Springfield, S. and Reynolds, D. (2000). Context issues within school effectiveness research. In Teddlie, C. and Reynolds, D. (eds.). *The International Handbook of School Effectiveness Research*. Falmer Press: London, pp. 160 - 185.

Theil, H. (1954). *Linear Aggregation of Economic Relations*. North-Holland: Amsterdam.

Thomas, S. and Mortimore, P. (1996). Comparison of value-added models for secondary-school effectiveness, *Research Papers in Education*, vol. 11, pp. 5 – 33.

Thomas, S., Sammons, P., Mortimore, P. and Smees, R. (1997). Differential secondary school effectiveness: comparing the performance of different pupil groups. *British Educational Research Journal*, vol. 23, pp. 451 – 469.

Thomas, S. (2001). Dimensions of secondary school effectiveness: comparative analyses over regions. *School Effectiveness and School Improvement*, vol. 12, part 4.

Trower, P. and Vincent, L. (1995). *The Value Added National Project, Technical Report: Secondary*. SCAA: London.

Tucker, P. and Stronge, J. (2005). *Linking Teacher Evaluation and Student Learning*. Association for Supervision and Curriculum Development: Virginia.

Vella, F. (1998). Estimating models with sample selection bias: a survey. *Journal of Human Resources*, vol. 33, pp. 127 – 169.

Werfhorst, H. van de, Sullivan, A. and Cheung, S. (2003). Social class, ability and choice of subject in secondary and tertiary education in Britain. *British Journal of Educational Research*, vol. 29, pp. 41 – 62.

Willms, J. (1986). Social class segregation and its relationship to pupils' examination results in Scotland. *American Sociological Review*, vol. 51, pp. 224 – 241.

Willms, J. (1987). *Comparing Schools in Their Examination Performance: Policy Questions and Data Requirements*. University of Edinburgh: Edinburgh.

Willms J. and Raudenbush, S. (1989). A longitudinal hierarchical linear model for estimating school effects and their stability. *Journal of Educational Measurement*, vol. 26, pp. 209 – 232.

Woodhouse, G. (1990). The need for pupil level data. In Fitz-Gibbon, C. (ed.). *Performance Indicators: a BERA Dialogue*. Multi-lingual Matters: Clevedon.

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

Zellner, A. (1966). On the aggregation problem: a new approach to a troublesome problem. In Fox, K. *et al* (eds), *Economic Models, Estimation and Risk Programming: Essays in Honor of Gerhard Tintner*. Springer-Verlag: New York.

Zuzovsky, R. and Aitkin, M. (1991). Curriculum change and science achievement in Israeli elementary schools. In Raudenbush, S. and Willms, J.D. (eds.). *Schools, Classrooms, and Pupils: International Studies of Schooling from a Multilevel Perspective*. Academic Press: London, pp. 25-36.