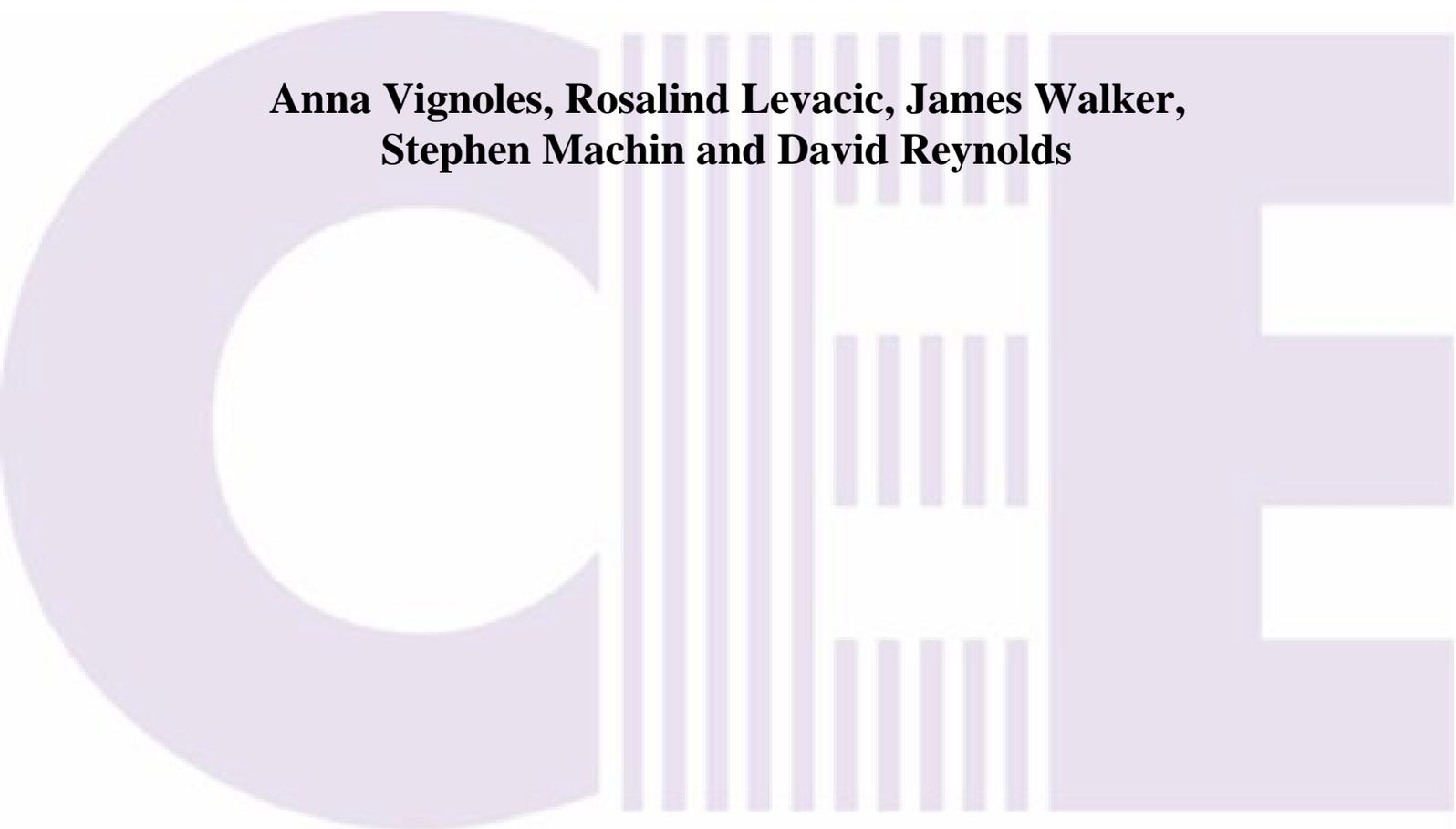


The Relationship Between Resource Allocation and Pupil Attainment: A Review

**Anna Vignoles, Rosalind Levacic, James Walker,
Stephen Machin and David Reynolds**



CENTRE FOR THE
ECONOMICS OF
EDUCATION

September 2000

Published by
Centre for the Economics of Education
London School of Economics and Political Science
Houghton Street
London WC2A 2AE

© Anna Vignoles, Rosalind Levacic, James Walker, Stephen Machin and David Reynolds

ISBN 0 7530 1433 5

Individual copy price: £5

The Centre for the Economics of Education is an independent research centre funded by the Department of Education and Employment. The view expressed in this work are those of the authors and do not necessarily reflect the views of the Department of Education and Employment. All errors and omissions remain the authors.

This work has been published in the Department of Education and Employment Research Report 228 (ISBN 1 84185 362 3) in September 2000. The original report remains Crown Copyright 2000 and was published with the permission of DfEE on behalf of the Controller of Her Majesty's Stationery Office.

EXECUTIVE SUMMARY

1. This report examines the impact of measurable resource inputs on primary and secondary school students' achievement. The objectives of this report are to summarise the findings from the educational production function literature, to identify the many methodological difficulties in this field, and to provide an outline for a high quality research project.
2. Section 1 concentrates on the methodological techniques most commonly used in the field – regression analysis¹ and stochastic frontier/data envelopment analysis - evaluating each technique and commenting on the advantages and disadvantages associated with each method.
3. Section 2 reviews a 'high quality' subset of the recent (post-Hanushek, 1997) international educational production function literature. This section has two aims. First and foremost, and in keeping with the methodological focus of the report, to evaluate how effectively the methodological issues have been addressed in the predominantly US centred literature. Second, to report the empirical results from this literature.
4. In Section 2 it is established that recent international research, in contrast to the widely held view that inputs have no systematic impact on student outcomes, has generally been able to establish a link between certain education inputs (particularly expenditure and teacher characteristics) and enhanced student outcomes.
5. Sections 3 and 4 present a review of the UK literature. Reflecting our interest in UK policy, and the limited amount of high quality research in the UK, the criteria for inclusion in this section are weaker than for Section 2 and consequently the review is more broad ranging. The review of the UK research is divided into two parts. The first (Section 3), reviews the production function based literature, while the second (Section 4) examines studies of cost effectiveness of school provision.
6. The UK literature review shows that, while the link between educational expenditure and outcomes is not proven, some real resources seem to have an impact on outcomes. For example, several studies found a correctly signed and statistically significant relationship between the school level pupil teacher ratio and outcomes. However, there is almost no UK evidence that smaller class size leads to better outcomes. In general, the UK literature is limited by the fact that many possible interactions between the various school inputs and resource variables under consideration have not been fully investigated. Hence, whilst in the UK school type appears particularly important in explaining examination performance; it is not clear to what extent this is due to the peer group effect, to better resourcing or better teaching quality in grammar, independent and single sex schools. More work is needed to fill this gap in the literature.
7. Section 5 draws on the methodological and conceptual issues identified in Section 1, and the empirical evidence (Sections 2, 3 and 4) to provide guidance to policy-makers about an 'ideal' research programme. In particular, the section discusses the data requirements to undertake such a study.

¹ A glossary of terms is provided at the end of this report.

8. Section 5 highlights the following:
 - The need for testing based upon theoretical modelling. The lack of systematic testing of existing theoretical frameworks has often led to incoherence and repetition in the literature.
 - The importance of explicitly addressing methodological difficulties in order to obtain plausible results. In particular, different methods (*e.g.* Instrumental Variables, Random Experiments) are needed to deal with the most important methodological problem, the potential endogeneity of school resourcing levels.
 - The crucial importance of having high quality data. Specifically, one needs pupil level data, with information on students' prior achievements and family background, supplemented with good school information and high quality resource information.

9. Section 6 draws conclusions and implications for policy. The most important conclusion is that further work is needed to investigate the link between resource levels and outcomes. Whilst we do not provide a full evaluation of the quality of all the possible UK data sets that might be used to research this area, we do make the case that the available UK data is insufficient to carry out a high quality study that would overcome most of the methodological problems identified in this review. However, once the Common Basic Data Set (CBDS) comes fully on line, it may meet many of the methodological criteria we have identified in respect of performance data. However, CBDS data needs to be linked to sound and consistent financial data in order to be used for the analytical purposes identified in this review. Links also need to be established between the CBDS and teacher databases and other sources of information on pupils' neighbourhoods and local environments. It is of primary importance that good financial data, based on comparable school accounting practices, are collected in the future. It is suggested that a 2-3 year longitudinal survey would meet policy needs in the medium term, *i.e.* before the CBDS comes on line.

The Relationship Between Resource Allocation and Pupil Attainment: A Review

**Anna Vignoles, Rosalind Levacic, James Walker,
Stephen Machin and David Reynolds**

Introduction	1
1. Methodological Issues	2
1.1 Regression Models	3
1.2 Educational production frontier models	12
2. International Evidence	16
2.1 Reviewing the reviews	17
2.2 Review of the recent international literature by input	20
2.3 Policy interventions	32
2.4 Concluding comments	35
3. UK Education Production Function Studies	35
3.1 LEA-level education production function studies	37
3.2 School level education production function studies	41
4. UK Cost Effectiveness Studies of School Provision	49
4.1 The cost-effectiveness of A/AS level provision	49
4.2 Educational interventions	54
4.3 OFSTED's qualitative assessment of schools' efficiency	59
4.4 UK education production function and cost-effectiveness	
Research: conclusions	63
5. The Ideal Research Project	65
5.1 The need for an ideal research project	65
5.2 Methodological considerations	66
5.3 Key variables required for a high quality research programme	67
5.4 Data issues	71
6. Conclusion and Implications for Policy	74
Glossary of Key Terms	76
References	79

Acknowledgements

All of the authors are members of the Centre for the Economics of Education. Anna Vignoles is also a member of the Centre for Economic Performance. Rosalind Levacic is at the Institute of Education. James Walker is at the London School of Economics. Stephen Machin is a member of the Centre for Economic Performance, London School of Economics and at University College London. David Reynolds is in the Education Department at the University of Newcastle upon Tyne.

The Centre for the Economics of Education is an independent research centre funded by the Department of Education and Employment. The view expressed in this work are those of the authors and do not necessarily reflect the views of the Department of Education and Employment. All errors and omissions remain the authors.

This work has been published in Department of Education and Employment Research Report 228 (ISBN 1 84185 362 3) in September 2000. The original report remains Crown Copyright 2000 and was published with the permission of DfEE on behalf of the Controller of Her Majesty's Stationery Office.

The Relationship Between Resource Allocation and Pupil Attainment: A Review

**Anna Vignoles, Rosalind Levacic, James Walker,
Stephen Machin and David Reynolds**

Introduction

Trying to accurately ascertain the direction and magnitude of any links between school resources and the educational attainment of pupils and their schools is an important research question that has potentially significant ramifications for education policy. There has been considerable research addressing this question and the goal of this review is to accurately appraise the key findings and methodological shortcomings of this work, with the intention of providing a pointer to the route that future research should follow.

The existing literature is not uncontroversial in the sense that there are clear differences in the findings emerging from studies that adopt different methodologies. For example, Hanushek's (1997a) review of the education production function field has suggested that there exists no systematic and statistically significant relationship between the level of educational resourcing (generally measured by expenditure per pupil and the average pupil teacher ratio) and pupils' educational outcomes. However, not all the available evidence supports this conclusion and there are a number of conceptual and methodological difficulties that plague the literature. Furthermore, even if there is currently no clear positive relationship between expenditure on schooling and pupil outcomes, this probably raises more questions than it answers. In particular, it suggests that researchers need to explore further issues relating to the way in which resources are allocated within schools, rather than simply looking at the relationship between aggregate expenditure per pupil and outcomes.

The primary purpose of this review is therefore to identify the central conceptual and empirical issues as regards the international and UK educational production function literature, focusing largely on the impact of educational expenditure and resource mix. This review is not intended to be exhaustive, rather the focus is on providing a methodological critique of this field of research which can guide policy-makers and point to the most appropriate directions for future research in the UK context. Although the parallel school effectiveness research field is referred to, the review is restricted to the educational production function literature, for two major reasons. First, it is the latter that has been most concerned with the impact of resources on outcomes and second, other excellent and comprehensive reviews of the SER literature already exist (Teddlie and Reynolds, 2000).

The report starts with a discussion of the key methodological issues in this field, identifying the most significant conceptual and empirical problems (Section 1). The methodological section attempts to respond, at least in part, to the recommendations of David Mayston and David Jesson (DfEE, 1999) concerning the need for more research into the appropriate techniques to model the relationship between resourcing and educational performance. In particular the use of the following techniques is discussed: linear regression (and value-added), multi-level modelling and frontier estimation models, including data

envelopment analysis².

In Section 2 we provide a brief overview of a selection of ‘high quality’ international studies³ from the educational production function literature. This international evidence is restricted to high quality surveys and key published work which has examined the following inputs: expenditure per pupil; teacher-pupil ratio; class size; teacher quality (education, ability, experience and salary); non-teaching expenditure and the following outputs: educational attainment (such as exam performance) and cognitive attainment.

Sections 3 and 4 examine the UK evidence. To ensure that the review is more extensive in its coverage of the UK evidence weaker criteria are used for the inclusion of studies. In particular, some unpublished work is included. The core objectives of the UK review are to identify pivotal findings from the UK research, as well as identify the major limitations of, and gaps in, the UK empirical evidence.

Section 5 relates the findings to possible directions for future UK research. In particular suggestions are presented for an ‘ideal’ research project, which take account of the methodological comments made in Section 1. Although the review does not go over ground already covered by Mayston and Jesson (DfEE, 1999), further strong support is provided for the case they made in favour of the proposed Common Basic Data Set. Some preliminary suggestions for particular data that would be needed to enable educational researchers to carry out top quality research in this area are also given.

1. Methodological Issues

To understand the methodological controversies surrounding this field of research, a historical view of the research literature is required. Several early and highly influential studies in this field appeared to show that schools hardly had any impact on the learning outcomes of children (in terms of their academic attainment, attendance and behaviour). First, in the US, the Coleman Congressional Report of 1966 (Coleman *et al*, 1966) was charged with looking at the relative educational opportunities for different racial groups in the US. Although the report did highlight the very different levels of school resources that were being allocated to African-Americans and white Americans, its results suggested that schools did not in fact make a great deal of difference to student outcomes. Family circumstances, ability and socio-economic background were found to be far more important. Their findings were partially challenged by the results of a major UK report entitled ‘Fifteen Thousand Hours’ (Rutter *et al*, 1979). This study showed that schools have a modest but significant impact on the learning outcomes of children, though again family circumstances, socio-economic background and individual ability matter far more than schooling. Since these two reports, this important debate has continued apace and there has certainly been a huge amount of research effort in this field. However, the results still appear to be ambiguous (Hanushek, 1986, 1996). This section highlights the reasons why this might be so.

There are three primary methodological problems associated with the Coleman and Rutter *et al* reports, and indeed the bulk of empirical investigation in this literature. First, there is a lack of an established ‘consensus’ theory from which appropriate models can be constructed. Secondly, the research that has been undertaken suffers from a number of technical/empirical problems. Lastly, again from an empirical perspective, poor data is a recurrent theme in this literature. These problems cannot be easily separated. Therefore a pragmatic view is taken which simply highlights how these difficulties relate to the empirical

² A full list of the techniques that have been used in this field is contained in Mayston and Jesson (1999, p.20). We do not explicitly consider the more accounting based methods or those techniques that do not allow fully conditional relationships to be modelled, *e.g.* ratio analysis.

³ Much of the influential literature in this field comes from the United States.

findings. Hence this section discusses two quite different methodological approaches in turn, starting with regression models (Ordinary Least Square (OLS) and related variants), followed by frontier estimation models (including Data Envelopment Analysis (DEA)). Each method's particular empirical and conceptual problems are described in some detail but it is important to note that many of the data related issues discussed initially in the context of regression models apply equally to frontier estimation models. To avoid repetition however, such issues are only discussed once.

To provide some background for the reader this review first considers some theory. Much of the empirical literature is based, explicitly or implicitly, on the theoretical concept of an educational production function. Using terminology from Cooper and Cohn (1997), an educational production set can be described by the following equation

$$F(y, x) \leq C \tag{1.1}$$

where y is a vector of educational outputs, and x a vector of inputs. C is a positive scalar, while F represents the educational technology which transforms x into y . Equation (1.1) describes the combinations of inputs and outputs that are technically possible, *i.e.* the production set. The maximum level of outputs for a given level of inputs is called the educational production function or frontier and represents the set of technically efficient solutions.

The production function approach operates by assuming that a variety of inputs (such as family background, educational resources and initial abilities of the child) are transformed by the school into a range of outputs, such as standardised test scores and examination results. Within this broad framework, there are two main methods of empirically estimating such relationships. The first is to utilise regression techniques. These tend to be parametric methods⁴ used to estimate 'average statistical behaviour' (Cooper and Cohn, 1997; Mayston and Jesson, DfEE, 1999). Regression analysis has been used to ascertain whether schools with higher resource levels also have higher performance levels, in relation to the *average* performance of all schools. These regression models generally also require the user to specify a particular relationship between the chosen inputs and the outputs of interest (Thanassoulis, 1993). The second main approach is frontier estimation. Frontier estimation can be either parametric (stochastic frontier regressions which specify the functional form of the stochastic production function) or non-parametric (Data Envelopment Analysis, DEA). These approaches evaluate the performance of schools in relation to the production frontier. The purpose of this approach is to identify those schools which have the best possible outcomes, for a given level of inputs, and which are therefore on the frontier of the educational production set.

1.1 Regression models

This Section starts by considering the most commonly used methodological approach, the broad family of regression models. In most regression analysis, a single educational output of interest is regressed against various explanatory variables or 'inputs'. The assumption is that greater quantities of inputs will, via some usually unspecified 'black box' educational technology, translate into higher output. This is certainly the method most commonly used to investigate the impact of school resources, such as expenditure, on learning outcomes.

The earliest studies that used regression models often lacked specific data on school characteristics, such as pupil-teacher ratios, and therefore only attempted to measure the total magnitude of any 'schooling effect'. This involved identifying the net effect on student

⁴ By parametric we mean requiring assumptions about the functional form of the mapping of outputs to inputs and the error process for that formulation. Of course there are non-parametric regression methods that do not impose such assumptions but, to our knowledge, these have not been used in the school resources literature.

outcomes of attending a particularly ‘good’ or ‘bad’ school. This was modelled by incorporating a set of dummy variables to measure the separate effect of each school in a standard regression of outcomes (such as standardised test scores, school completion rates, truancy rates) on various explanatory background variables, such as family background. Generally, the results convincingly indicated that schools do make a difference. For example, Creemers and Reezigt (1996) concluded, based on UK and other Western hemisphere data, that around 10-20% of the variance in student achievement is attributable to school factors. Reynolds *et al* (1996) put the figure at nearer 8-12% and found primary school effects to be larger than secondary school effects.

Yet when researchers progressed from identifying school effects to analysing the effect of particular inputs, such as family background and educational expenditure, the results were disappointing. Using these regression techniques, researchers have had great difficulty in identifying which factors, particularly in terms of resources, make some schools more effective than others. So what have been the factors or inputs of greatest interest to researchers? As discussed at length below, empirical work in this field needs clearer guidance as to the inputs and outputs of *theoretical* importance. In practice, the outcome variables of interest have generally been standardised test scores, exam results and staying on rates. As shown in the equation below, the pupils’ outcomes (O_{is}) are regressed against various school quality variables S_s (that are generally measured at the school or school district/Local Education Authority (LEA) level), and a set of other background variables (X_i). Formally,

$$O_i = f(S_s, X_i) \quad (1.2)$$

The most commonly used proxies for school quality are the pupil-teacher ratio and expenditure per pupil. Other possible indicators of school quality include school size and type of school, *e.g.* private or grammar. Although, as has already been stated, much of the empirical literature does not appeal to an established theoretical base for such models, the general hypothesis is that greater school resources will have a positive effect on pupils’ learning outcomes⁵. In this context, the regression methodology has a number of theoretical and empirical problems, which are discussed below in approximate order of importance.

1.1.1 The possible endogeneity of school quality⁶

The most serious problem in the literature as a whole is the potential endogeneity of educational quality. Certainly parents who send their children to private schools can choose the quality of that schooling. Even parents who send their children to state maintained schools may be able to get their children into a ‘better’ school, for example by buying a house in a better neighbourhood. If this occurs, school quality will be positively correlated with the wealth and social advantage of children’s families. If wealth and social advantage impact on students’ learning irrespective of school quality, then some of the apparent gain from additional school quality will in fact be a ‘return’ to pupils’ socio-economic background.

The literature has also identified another endogeneity problem that is likely to cause a bias in the opposite direction. Some school funding systems operate compensatory resourcing arrangements, whereby greater resources are allocated to poorer areas or schools or applied to weaker students. The aim of such funding is generally to compensate for the effects of disadvantage on performance. In making this connection between likely performance and funding levels, a relationship is created which will, to some extent, mask the

⁵ This issue applies equally to research that has used other estimation methods such as DEA.

⁶ In fact the general endogeneity problem applies to the literature as a whole, including research that has used other estimation methods.

analysis of the relationship between school inputs and pupil attainments (Burtless, 1996). If the link between socio-economic characteristics and funding is not fully controlled for, a model of attainment will tend to generate a spurious negative correlation between school resources and achievement. The system of educational funding in the UK includes such a compensatory measure. In particular, the national funding system allocates resources to local education authorities on the basis of pupil numbers in various age bands, weighted by factors to reflect social need in the authority and also, in some cases, higher cost indices. LEAs own school funding formulae then pass these funds to schools in ways which reflect need in varying degrees. Schools themselves are then free to allocate resources to students, for example by differential teaching group sizes. Indeed Goldstein and Blatchford (1998) argue that, with respect to endogeneity in UK observational studies, the relationship between class size and pupil value added attainment is affected by schools placing lower attaining children in smaller classes and deploying better teachers in larger classes. As a result it is likely that researchers will find spurious positive and significant relationships between class size and outcomes.

The endogeneity problem is at the root of a number of theoretical and empirical critiques of the findings in this field (e.g. Mayston, 1996). For example, Mayston talks of the need to model the demand side of the educational market, *i.e.* the behaviour of LEAs and schools. He argues that most educational production function models focus exclusively on the supply side, *i.e.* the simple link between inputs and outputs, and he shows that the level of resources experienced by a child will be endogenously determined if schools undertake optimising behaviour. To illustrate, assume that a school is given a fixed budget. Assume also that the school knows that the same level of resource inputs has a very different effect on a child's attainment, depending on the socio-economic background and prior attainment of the child. The school will therefore allocate their fixed amount of resources among their students, taking this fact into account. In other words, the school will systematically allocate resources to each child, such that the learning output of the whole school is maximised. This is really a re-statement of the classic endogeneity problem that is associated with single equation regression models. This criticism suggests that, contrary to Hanushek's conclusion (1996) that school resources do not impact greatly on pupil outcomes because schools are inefficient, resources do not appear to impact on outcomes because schools are efficiently optimising their use of these scarce resources. There are a number of possible empirical solutions to this problem that are examined below.

1.1.1a Value-added models

One way of overcoming some of these endogeneity issues, and isolating the effect of school quality, is by estimating 'value-added' models⁷ which allow for the pupil's initial ability and socio-economic background, their school's socio-economic composition (for example the percentage of students eligible for free school meals), the gender of the pupil and their ethnicity. Why may these improve upon the basic regression approach? It is clear that an individual's initial ability and attainment prior to starting at that school are obviously important as they allow for the fact that some children are intrinsically more able than others, some have already had better schooling than their peers, and some have experienced greater parental inputs before starting at a particular school. In other words, the inclusion of these variables effectively 'levels the playing field' at time of school entry. Clearly the impact of the child's subsequent schooling must be measured separately from these other factors. To allow for this the value-added method adopts one of two approaches. First, it uses a dependent variable measured as the *change* in a student's test results over a particular period of schooling. Second it includes the child's initial test results (prior to starting at the school

⁷ A value-added approach can also be used with frontier estimation techniques. See next section for a fuller discussion.

or at the beginning of a particular educational intervention) as an explanatory variable on the right hand side of equation (1.1).

The value-added method is certainly a vast improvement on single equation educational production function models that do not even take into account the child's attainment on entering school. Indeed, it is now well established that if teachers, schools or LEAs are to be rewarded (or held accountable) for their performance, the use of a value-added approach is essential. However, value-added analysis is not in itself sufficient to overcome the endogeneity problem. Even allowing for the child's initial attainment, it may still be the case that the effect of the resourcing that the child actually experiences is systematically related to their family background and prior attainment, and that most value-added models are not constructed to deal with this problem.

1.1.1b Simultaneous equation models or instrumental variables (IV)

A limited amount of research has attempted to tackle the endogeneity problem by using simultaneous equation models that explicitly model the resource allocation between students and between schools (see Mayston, 1996) for a summary and Mayston and Jesson (DfEE, 1999). In the UK context, if such models are to be developed, a better understanding of the resource allocation process within schools and LEAs is needed. Researchers need to identify and model the determinants of school resources per pupil, as a first stage in a full structural model of the effects of school inputs on learning outcomes.

In the simultaneous equations approach the structural relationships are made explicit and form a system of equations that permit the untangling of the structural associations between multiple inputs and multiple outputs. In particular, if properly specified, the simultaneous equations approach can enable researchers to identify both the impact of the various factors that influence the level and mix of educational resourcing a child receives (*e.g.* through the Standard Spending Assessment, school funding formula and different class allocation methods), and the effect of these resource levels and mix on educational outcomes. The main drawback of this approach is the identification problem. To solve the equation system, the researcher must find a factor which influences the amount of resourcing a child receives without also influencing that child's educational performance directly. Furthermore, the more equations there are in the system (as would be the case if the researcher was attempting to model multiple inputs and outputs), the more identification restrictions one requires and therefore the more difficult the identification problem is to solve.

Probably a more practical approach than using a simultaneous equations model is to find an instrument (or instruments) for the potentially endogenous school resources variable(s), rather than specifying the full set of equations complete with feedback mechanisms between different dependent variables. Of course this is the same thing as the identification issue discussed above, as a variable needs to be found that exerts no direct influence on outputs and only works indirectly through its role as a predictor of resources. Conceptually and theoretically this is difficult. However, the Instrumental Variables (IV) method provides perhaps the most fruitful prospect for good quality research in this field. Indeed, existing results from IV estimation are encouraging. Specifically a number of influential papers that have used IV have found positive effects from school inputs (Akerhielm, 1995; Angrist and Pischke, 1999; Figlio, 1997). However, it is not universally the case that good quality studies that use IV find positive effects from school inputs (Hoxby, 1998). There are also a number of difficulties with the IV approach.

IV methods, like simultaneous equation models, have the problem of identifying factors (*i.e.* instruments) which influence the allocation of school resources among students but which do not in themselves influence learning outcomes. Certainly researchers have come up with a number of innovative ways of identifying natural random variation in school inputs. For example, Figlio (1997a) uses the tax revenue raising limits that have periodically

been imposed in certain US states (akin to the capping of local councils in the UK) to identify 'random' changes in educational expenditure. Hoxby (1998) uses natural changes in population (birth rate) and externally imposed class size limits to identify random variation in class size. In other words, sudden increases (decreases) in the birth rate, combined with legal limits on class size, mean that part of the variation in class sizes in a given year is due to these external factors, rather than the characteristics of the children in these classes. This exogenous variation can then be used to identify the effect of class size on outcomes. The search for appropriate instrumental variables, and the development of theory to suggest these, is clearly an avenue for future research.

A further problem with using IV models is that they can only identify the effect of a change in school inputs for a particular sub-set of the pupil population where there has been a variation in school inputs. For example, unexpected changes in the birth rate/population are likely to have a bigger random impact on class sizes in smaller schools and school districts, where administrators cannot smooth out the effect of sudden changes in enrolments. These small schools and districts are not representative of all schools and districts so the results from studies which use this type of instrument must be interpreted with great care and the testing of the statistical validity of instruments is essential (Bound *et al.*, 1995). Equally studies that have used rules about maximum class size as instruments are actually measuring the effect of random change in class sizes around the maximum possible class size (40 in the case of Angrist and Lavy, 1999). It is quite possible that changes in class size from 40 to 41 may have a significant effect, even if changes in class size from 30 to 31 (the range more relevant to the UK) do not. Furthermore, particular students may be clustered in the classes most affected by class size rules. If the characteristics of the target population (children in classes affected by class size rules at the margin) are not representative of the total population, results may be biased.

1.1.1c Randomised experiments

One virtually guaranteed way to overcome the endogeneity issue is to use randomised or experimental data, of the kind employed by Krueger (1999). The "STAR" class size experiment in the US exemplifies the use of experiments in social science, and in education specifically. The advantages of this methodology are clear. With a carefully constructed experiment, the potential for endogeneity is limited (although probably not altogether eliminated). In an ideal world all educational interventions might be subject to randomised trials prior to their introduction.

However, there are a few disadvantages associated with random experiments. From a practical point of view they are costly and may raise ethical issues. For example, parents may feel they do not want their children to be subjected to 'experimentation'. Hoxby (1998) also argues that there are still a few methodological problems with random experiments. Random experiments tend to be extremely rare and therefore their results are cited often in the literature and extrapolated to other, sometimes very different, institutional settings. For example, the results from the US class size experiment can only be applied to the UK with great caution, given the institutional differences between the two education systems.

The most serious criticism of random experiments is that the individuals (students, parents, and school staff) involved in the experiment are generally aware of it. This may lead to "Hawthorne" effects, whereby students perform better just because they are the subject of an experiment, rather than due to the educational intervention itself. In particular, as the experiment may lead to a particular policy recommendation (smaller classes) all involved have an incentive for the experiment to 'work'. Equally individuals involved in the experiment may tend to subvert the random nature of the experimental design (Hoxby, 1998). For example, parents may lobby for their children to be put in the smaller classes. Teachers may put certain students, who they feel would benefit most, in the smaller classes. All this is

really to say that even random experiments may not totally eliminate the endogeneity problem. Efforts to find natural random variation in school inputs (to be used as instruments) may therefore be, not only more feasible, but also advantageous from a methodological point of view.

1.1.2 The problem of model specification

As has already been emphasised, one of the difficulties in this field of research is the lack of an established theoretical model of how school resources might impact on educational outcomes⁸. For example, Blatchford and Mortimore (1994) stress the importance of clearly specifying a model, derived from teaching and learning processes, to explain the relationship between class size and performance. Such models would tend to stress the importance of educational processes, as well as inputs. Generally however, a ‘black box’ approach is taken, whereby ‘resources’ are simply applied to a pupil, school or school district and it is hypothesised that this will in itself generate better learning outcomes. Certainly in studies which focus on the impact of school resourcing, the processes and resource mix needed to ensure that additional inputs result in better outcomes are generally ignored. There are however, a number of researchers whose theoretical models may provide guidance as to the correct specification for empirical models (Carroll, 1963; Walberg, 1984; Creemers, 1994; Reezigt *et al*, 1999). Most of these models attempt to answer important theoretical questions in this field such as: what are the important factors that influence learning outcomes; what are the important inter-relationships; what factors are important at what level, *i.e.* at the pupil, classroom and school levels; what are the important cross level interactions? Yet despite this work, much of the empirical literature has not referenced such models, so that a coherent body of empirical evidence, that may be systematically tested, has not been built up.

To illustrate this point, consider Carroll’s model of student learning (Carroll, 1963). This model, on which many later models (Creemers, 1994); Scheerens, 1997) were based, hypothesises that a student’s learning rate depends on five factors; “aptitude, ability to understand instruction, perseverance, opportunity and the quality of instruction” (Creemers, Scheerens and Reynolds in Teddlie and Reynolds (2000, chapter 11, p2). Furthermore, Carroll (1963) suggests that *time*, in terms of the time-spent learning and the time needed by a particular student to learn, are crucially important determinants of achievement. The five factors identified by Creemers (1994) can be broken down into pupil factors (aptitudes *etc.*) and classroom or school level factors (for example, quality of instruction). Although this is just a simple illustration, it shows how a theoretical model can inform empirical research in this field.

Certainly, as a result of the lack of a theoretical basis for much of the empirical work in this field, the problem of omitted variable bias is pervasive. For instance, much school quality research implicitly assumes that expenditure per pupil, the pupil-teacher ratio and certain measures of teacher quality are adequate proxies for the quality of a school or school system. Yet, as the mixed evidence from both the UK and US suggests, education is a highly complex process. There may be a multitude of important factors that are generally omitted from studies of school quality, possibly biasing the estimated effects of the resource measures that are included. For example, omitting measures of the time parents spend with the child may bias results, as will including variables that may be inappropriate and collinear with the school inputs of interest. Dewey *et al* (2000), for example, provide some indication that studies that include family income or socio-economic background, instead of parental time inputs, are less likely to find positive effects from school inputs. The suggestion is that family income is actually a poor measure of parental inputs and, since it is correlated with school inputs, may cause a downward bias on the school input variables. Indeed, Hanushek

⁸ This problem is also acute for research that uses other estimation techniques such as DEA.

(1986) makes the more general point that the educational inputs used in empirical work tend to reflect data availability, rather than any particular theory about the determinants of school quality.

It is also true that even if researchers base their work on a theoretical model and also have access to good quality data on school resourcing levels, data sets often lack adequate measures of family background, peer effects and community environment. It is widely believed that these factors may be as important as the school environment (Coleman *et al*, 1966; Haveman and Wolfe, 1995; Gregg and Machin, 1999). If this is so, and if the omitted variables are correlated with the measures of school quality used, then the coefficients on the school quality variables will be biased (Altonji and Dunn, 1996 and Behrman *et al*, 1996). The solution to the general problem of omitted variable bias is to make better recourse to the theoretical models in the field and, where possible, to use richer data.

1.1.2a Multilevel models and data aggregation

Continuing with the theme of model specification issues, Goldstein (1987, 1995) has argued that there are statistical problems associated with modelling pupil outcomes and not allowing for the hierarchical nature of the data, *i.e.* the clustering of students in classrooms, classes within schools and schools within LEAs. Intuitively, the problem is that some of the difference in pupil performance may be attributable to factors linked to the child (family background) and yet some of the difference may be due primarily to the class or school the student attends. Hence, all the children in the same class/school will therefore share a positive (or negative) effect from being in that class/school. This class/school effect may in turn be linked to good/bad management or equally to higher class/school resourcing levels. However, ignoring the ‘class or school effect’ in statistical estimation will lead to bias. To model this clustering, a variance components or multilevel model is needed. Assume that there are just two ‘levels’ in the data, the pupil level and the school level. The achievement of student i at school s (O_{is}), is given by

$$O_{is} = \alpha + \gamma_s + \epsilon_{is} \quad (1.3)$$

where the variation in pupils’ outcomes depends on the variation between schools, γ_s , and also on the variation between individuals ϵ_{is} ⁹ and where α is a constant term. The second stage of the analysis then allows the researcher to explain the variation between schools that is linked to school resourcing or other factors, as well as the variation between pupils that is related to resource inputs and other factors. So for example, the school pupil-teacher ratio might be used to explain the variation between schools, whilst the pupil’s family background and their actual class size could be used to explain the variation between individuals.

The complexities of multi-level modelling, and a more in-depth discussion of the statistical properties of such models, can be found in Goldstein (1987, 1995). However, it is important to stress that the prime advantage of the multi-level modelling approach is that it recognises the inherently hierarchical structure of an education system. This enables researchers to comment on the factors influencing performance at the different levels within the education system. Generally results show that multi-level models and OLS models yield similar results when there is only a low correlation between pupil outcomes within classes, schools or LEAs, *i.e.* there is low intra-unit correlation. It can be shown that in a two level model (*e.g.* pupil and school), with just one explanatory variable, as the intra-unit correlation increases, the OLS estimator will increasingly underestimate the true standard error. This will tend to lead to false rejections of the null hypothesis that the coefficient on the

⁹ These random components have an expected value of zero, are assumed to be uncorrelated and have variances σ_s^2 and σ_{is}^2 respectively.

explanatory variable is equal to zero. Hence not taking into account the hierarchical structure of the data in this instance may cause researchers to find a positive relationship between the explanatory factor and pupil outcomes, where none really exists.

There is also considerable evidence that a lot of the research in this field is affected by the closely related problem of ‘aggregation bias’ (Summers and Wolfe, 1977; Betts, 1995; Hanushek *et al*, 1996; Heckman *et al*, 1996a; Grogger, 1996). Although ideally researchers need to know the actual school resourcing experienced by the child, they have often had to rely on very aggregated measures of school quality, *e.g.* state level data in the US, which has led to biased estimates¹⁰. Hanushek’s surveys (1996 and 1997a) show that studies that use aggregated data are more likely to show a significant relationship between school quality and outcomes. For example, of the 77 estimates that measure the impact of pupil-teacher ratios using classroom level data, only 12% found that lower pupil-teacher ratios were associated with better student outcomes. On the other hand, of the 11 estimates that use aggregated state level data, 64% suggested that lower pupil-teacher ratios would have a positive effect on outcomes. The reason for this so called aggregation bias is that as the level of data aggregation increases, so too does the effect of bias from omitted variables. As has been discussed, many studies necessarily omit key variables, such as those measuring the community environment (Hanushek *et al*, 1996), which generates biased results. Although when many relevant variables are omitted the direction of the bias is difficult to predict, the seriousness of the problem is increased because ‘... as data are aggregated to the level of the omitted variable (*e.g.* state average data are used when state level factors are left out), any bias must worsen.’ (Hanushek *et al*, 1996, p.88).

In summary, an ideal research programme in this field would take account of clustering within classes, schools and LEAs by using a multi-level approach - or other appropriate econometric methods that allow for higher level effects - and would also use pupil level data on the outcomes of interest¹¹.

1.1.2b Functional form

Another issue relating to model specification arises because the regression models applied in most studies assume a linear (or log linear) functional form. This linearity assumption implies that the effect of an additional unit of school inputs is the same both at very low and at very high initial levels of school inputs. Yet this linear functional form has been statistically rejected by the data in some instances (Figlio, 1999), suggesting it is overly restrictive because it fails to allow for non-linear effects. Indeed, Figlio (1999), using a translog functional form, and Eide and Showalter (1999), using quantile regression techniques, found evidence of non-linearity. It is difficult to determine whether functional form is a major issue, in terms of the wider empirical findings in the literature. Figlio (1999) for example, rejects the assumption of linearity in the effects of school inputs but equally finds that any positive effects from school inputs are still very small. However, high quality research in the future would clearly benefit from more rigorous statistical testing of the functional forms used.

¹⁰ Indeed some researchers have attempted to overcome the potential endogeneity of school quality by looking at differences in educational resources and outcomes at a more aggregated level, *i.e.* across school districts, LEAs or states. This method may avoid the endogeneity issue if one makes the assumption that school quality may be endogenous to the individual but not for the state as a whole. In other words, whilst more privileged children may get better quality schooling within a state, it is assumed that richer states will not necessarily have better quality school systems. However, this assumption seems doubtful and, given the problem of aggregated data bias, pupil level data is widely seen as far superior for this kind of analysis.

¹¹ Although in terms of the explanatory variables, class, school and LEA-level data would be needed to supplement the pupil level data.

1.1.3 Interactions

Another way to attempt to overcome some of these endogeneity difficulties, in conjunction with the value added approach, is to estimate more saturated models, specifically using the 'proxy method' (Dearden *et al*, 1997) of controlling fully for factors that might influence the resourcing level experienced by a child. This approach is like the value-added approach, which tries to 'level the playing field' at time of school entry. It also stresses that researchers need to take into account important interactions between the school inputs and other variables. Certainly much of the empirical work in this field has assumed that school resources have the same effect on learning outcomes for all students, regardless of the family background or initial ability of the pupil. Yet studies that have examined the impact of school resources fails to take account of differing levels of ability or of differing socio-economic background have often found some significant results (Dearden *et al*, 1997; Figlio, 1999, and Wright *et al*, 1997). Further work needs to build on this approach to answer many more complex questions such as, do smaller class sizes benefit lower or higher ability children?

Once again it should be noted that clear theoretical guidance as to the possible interactions between different inputs is sorely lacking. Production function theory provides a useful framework, but further work is needed to develop and test educational theories¹² that might point to important interactions between resources and other inputs. For example, how does pupil aptitude interact with the quality of instruction to impact on outcomes (Carroll, 1963). Such theories, and the implied interactions between the various inputs, might then be systematically tested on different data sets. Only then will more definitive answers in this field be attainable.

1.1.4 Dirty data and errors in measurement¹³

Errors in measurement of the school quality variables may bias results downward (Behrman and Birdsall, 1983). If the variables are measured with error, then the coefficients may in fact underestimate the true effect of school resources on outcomes. Hanushek *et al* (1996) suggests that data aggregation under these circumstances may be helpful and reduce the bias from errors in measurement. However, the fact that measurement error will tend to bias estimates of the effect of school inputs downwards reinforces the message that good quality data are essential.

1.1.5 Cohort issues

There may also be a cohort issue (Bound and Loeb, 1996). As the quantity of resources put into education has increased, the gain in learning due to these resources may have fallen. Hence a positive relationship between school inputs and outputs may only be observed for earlier cohorts (born prior to the 1950s). Indeed Hanushek (1997a) and others have suggested that if there is a positive but diminishing 'return' to school quality measures, this may mean that the large increases in school resourcing over this century have ensured that we are on the flat peak of the education production function. Thus further increases in school resourcing would be unlikely to have a positive impact on outcomes. Research from developing countries, where educational expenditure per pupil is considerably lower, may shed light on this issue.

¹³ See Reezigt *et al.*, (1999) for a summary.

¹³ Obviously these data problems apply to the frontier estimation techniques discussed below.

1.2 Educational production frontier models

Having highlighted some general problems that relate to this literature as a whole (endogeneity issues, dirty data *etc.*), and some limitations that pertain specifically to regression models, the discussion now moves on to frontier models. As has already been mentioned, frontier models do not avoid many of the problems discussed in Section 1.1, particularly those relating to the lack of theoretical models and the need for good quality data. However no attempt is made to repeat the arguments made above in this section.

In fact the distinction between the regression method and frontier models is somewhat misleading. There are a number of different ways of estimating the educational production frontier, including parametric techniques (stochastic frontier regression) and non-parametric methods (Data Envelopment Analysis)¹⁴. Various studies have compared the different frontier methodologies and some have found that the different methods yield similar results. Some have argued that frontier methods can be considered superior to standard OLS regression in this context (Ganley and Cubbin, 1992). This report takes a more critical view, highlighting both the advantages and disadvantages of frontier techniques.

The frontier approach tries to identify the performance of individual schools in relation to the educational production frontier, represented by equation (1.4):

$$F(x, y) = C \tag{1.4}$$

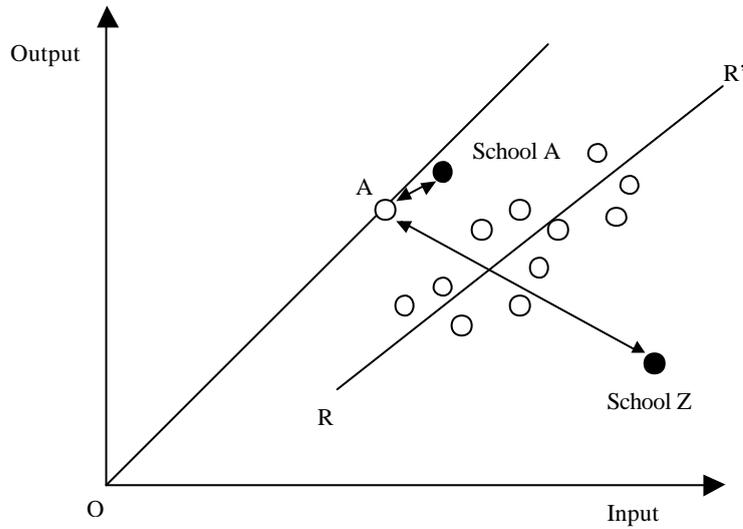
where y is a vector of educational outputs, and x a vector of inputs. C is a constant.

At first glance, frontier methods have a number of desirable features. First, by defining a technically efficient school (or set of schools) they may identify best practice behaviour (*i.e.* schools on the frontier). If measured accurately, this may be more useful to policy-makers than identifying differences from the mean, as in the standard regression approach. Furthermore, the frontier approach, by emphasising best practice, should avoid expectations being determined by average, rather than best, performance. The other attractive feature of frontier methods is that, by and large, the unit of observation is the school. Specifically, since frontier methods focus on the relative performance of schools, they are useful in providing estimates, which may be used to hold schools accountable for their performance and to ensure that they provide good value for money.

Graphically, the difference between the stochastic frontier/DEA techniques and regression analysis is depicted in Figure 1, in the case where there is a single output and input, and constant returns to scale is assumed. The segment RR' is a fitted line through the data points, as determined by regression analysis, while the distance from the efficient boundary (OA) is a measure of efficiency, as determined by DEA techniques. For example School A, which is located on the boundary is conceived as the most efficient school, while School Z, which is farthest from OA is the least efficient school.

¹⁴ More of this section is dedicated to the latter since only a small number of studies have used stochastic frontier regression.

Figure 1: Regression analysis and stochastic frontier/DEA



Having briefly considered the overall advantages of such methods, the drawbacks of each frontier estimation method are discussed separately, starting with stochastic frontier regression.

1.2.1 Stochastic frontier regression

A good example of the use of stochastic frontier regression is found in Cooper and Cohn (1997). Like all single equation regression models, they have to assume just one output (\hat{y}) but, unlike standard models, decompose the error term (e) into two components (v_i and u_i). Algebraically,

$$y = F(x, \mathbf{b}) + \mathbf{e} \quad (1.5)$$

where

$$\mathbf{e}_i = \mathbf{u}_i + v_i \quad (1.5.1)$$

Once again y and x are vectors of outputs and inputs respectively, and β is a vector of parameter estimates. Intuitively, v_i captures the stochastic noise term, which allows the frontier to shift between different schools for external reasons, thus yielding a stochastic frontier, while u_i is a non-positive error term that measures the technical inefficiency of the school, *i.e.* the distance from the frontier. In this specification, it is theoretically impossible for schools to perform *above* the educational production frontier, and hence u_i must be a non-positive term. Econometrically, Cooper and Cohn (1997) estimate a stochastic frontier by including the parameter λ , and estimate the equation using log likelihood. λ is given by

$$\lambda = \frac{\mathbf{s}_u}{\mathbf{s}_v} \quad (1.6)$$

The larger θ , the more dominant the inefficiency error term (\mathbf{s}_u), as compared to the noise error term.

An advantage of this form of frontier estimation is that it gives easily interpretable results, in the same way as the standard regression approach. Hence the frontier estimation results in Cooper and Cohn (1997) are contrasted with standard OLS results. The other advantage is that it is possible to test the statistical significance of the parameter values of various explanatory variables, which is obviously important in determining relations between the relevant variables, for example, whether class size significantly affects outcomes. However, because of the way in which it sets up a parametric production function and merely makes different assumptions about the error process, the frontier production approach also shares the problems associated with standard regression models. In particular, restrictive assumptions about the functional form of the model and the error terms are needed (Ruggiero, 1996). Perhaps as importantly, the method does not easily allow multiple outputs.

1.2.2 Data envelopment analysis

Data Envelopment Analysis has been used more extensively than stochastic frontier estimation in the education field.¹⁵ It is a non-parametric method of estimating the educational production frontier. The unit of analysis in DEA models tends to be schools or school districts. In simple terms, DEA estimates the performance of schools, relative to the educational production frontier, by identifying those schools on the frontier. The schools on the frontier are the ones that minimise their use of inputs for a given level of output, or conversely maximise their output for a given level of inputs. Hence DEA singles out the 'efficient' schools operating on the frontier and then measures how far all the other schools are from that frontier.

A mathematical representation of the optimisation problem based on the Kirjavainen and Loikkanen (1998) study, is as follows. Assume that there are n schools. School j produces the amount y_j of output r , using amount x_j of input i . Assume that both the inputs used and the outputs produced by each school are non-negative and that each school uses at least one input to produce one output. Assume the input weights are given by v_i ($i=1..m$) and output weights by u_r ($r=1..s$). In this simple case, the objective is to maximise the sum of the weighted outputs of school 0, subject to the sum of its weighted inputs being equal to 1. Thus,

$$\max w_o = \sum_{r=1}^5 u_r y_{r0} \quad (1.7)$$

$$s.t. \sum_{i=1}^m v_i x_{oi} = 1 \quad (1.7.1)$$

The optimisation problem is subject to the constraint that the sum of the weighted outputs of all the schools minus the weighted inputs of all the schools is less than or equal to zero, such that all schools are operating on or below the production frontier.

$$\sum_{r=1}^5 u_r y_{rj} - \sum_{i=1}^m v_i x_{ij} \leq 0 \quad (1.8)$$

$$j = 1..n : r = 1..s : i = 1..m : u_r, v_i \geq \epsilon$$

¹⁵ For an example of the early use of DEA and a comparison with standard regression techniques see Thanassoulis (1993).

This formulation allows each school to have its own weights, in terms of its inputs and outputs, and yields an efficiency score for school 0 of between 0 and 1¹⁶.

DEA therefore provides an efficiency score or index, sometimes called a Farrell efficiency index, for each school. Note that (1) the technique, by definition uses data aggregated to the level of the decision-making unit (generally the school or school district), and (2) the choice of inputs has varied from study to study, although some researchers have included measures of student intake, including socio-economic status, school characteristics, such as school size, school inputs, such as expenditure per pupil, and teacher inputs and characteristics, such as teacher education.

DEA analysis itself may highlight efficient and less efficient schools and also give an estimate of the efficiency gains if all schools were as efficient as the best performers in the sample. However, policy-makers are also interested in the effect of specific inputs, such as expenditure, and DEA does not provide a quantitative estimate of the impact of any particular input. Hence some studies (Bradley *et al*, 1999; Kirjavainen and Loikkanen, 1998) also explore the *determinants* of school efficiency in a second stage regression analysis. In this second stage the schools' efficiency scores are regressed¹⁷ against a number of factors that might explain the variation in efficiency between schools, *e.g.* the extent of competition between schools in the area, local environment and, of course, school resources. These second stage regressions can then provide evidence on the effect of specific inputs, in conjunction with the DEA analysis which gives guidance on the relative performance of schools.

The key advantage of the DEA method is that it can handle multiple inputs and outputs, as well as a value-added formulation.¹⁸ Furthermore, the model does not require the researcher to have information on the price of any inputs. In a public sector context, and particularly in education, decision-making units tend to generate multiple outputs and use multiple unpriced inputs. For instance, schools might produce both academic learning (measured by exam results) and citizenship (measured by propensity to vote) or socialised behaviour (turning up on time to school). They also use multiple inputs (teacher time and teacher quality), some of which are not easily priced. Furthermore, DEA is a non-parametric method that does not require restrictive assumptions about the functional form of the model, as is the case with the regression techniques used in the literature. Also, as has already been discussed, DEA identifies best practice or beacon schools (Mayston and Jesson, DfEE, 1999). Hence by estimating the relative efficiency of each school, it can help researchers answer questions such as: what would be the total gain in output if all schools were operating on the educational production frontier?

There are however a number of clear disadvantages with DEA. The first is a conceptual one. The DEA technique provides researchers and policy-makers with a measure of the relative efficiency of each school. However, this is merely a measure of the performance of that school relative to the best school in the data, rather than relative to any objective standard of technical efficiency. It is important that the results of DEA are therefore reported responsibly, without giving the impression that DEA provides an indication of the absolute efficiency of schools. It is also true that noisy data (*i.e.* that contaminated by measurement error), or a lack of detailed enough input variables, may cause misclassification errors in the school reported to be most efficient. In this regard regression modelling of the average (or use of quantile regressions that pick out specific percentiles of

¹⁶ The particular formulation described here assumes constant returns to scale. If variable returns to scale are assumed, the optimisation problem is marginally more complex, and is not presented here (see Ganley and Cubbin, 1992 and Kirjavainen and Loikkanen, 1998).

¹⁷ Normally using Tobit estimation to allow for the censoring problem.

¹⁸ By including students' socio-economic status and prior ability either into the original DEA model or the second (regression) stage of the analysis.

the distribution, but not the single highest value corresponding to maximum efficiency) may well be preferable.

Furthermore, from a technical perspective, the non-parametric nature of the estimation technique is also a serious disadvantage. DEA does not allow researchers to make statements about the statistical significance of the relationship between certain inputs and outputs, *e.g.* the effect of class size. When the two-stage approach is adopted where technical efficiency from the first stage is regressed on various characteristics, the possible drawbacks of the regression approach re-emerge. Moreover, there is a serious identification issue to do with this two-stage approach, namely that characteristics used to explain relative efficiency in the second stage are not allowed to enter the first stage. In most cases this is an unrealistic assumption to adopt.

Furthermore, Ruggiero (1996) suggests that, since DEA is non-stochastic, it is particularly sensitive to measurement error and variable selection. The former issue is likely to be a particular problem in relation to the measurement of educational expenditure. The latter point about variable selection is even more important. Although DEA may not impose a particular functional form on the model, it does require the researcher to choose the relevant inputs, and to decide at which stage of the analysis each input belongs (*i.e.* in the DEA or the second stage regressions). There is clear evidence that the results of any DEA analysis are sensitive to the timing of inclusion, given the inability to statistically test the model in the initial stage and the lack of coherent theoretical guidance to inform the choice of inputs used.

A last technical problem is that there is some empirical evidence that suggests that the results of DEA are sensitive to assumptions made about the returns to scale in education production (Kirjavainen and Loikkanen, 1998)¹⁹. As the DEA method is non-parametric, there are no statistical tests available to test such assumptions, and hence it is crucial that work in this area checks the robustness of the efficiency rankings generated by DEA to assumptions made about returns to scale, and, in the light of comments about measurement error, to outliers in the data.

This section ends by noting that many of the methodological issues discussed above relate to poor data. Mayston and Jesson (DfEE, 1999) have already made a persuasive case for a national database that would enable better links to be made between educational resourcing and performance. The methodological points raised here need to inform the final decisions regarding the exact nature of any such database, as discussed in Section 4.

2. International Evidence

It has been three years since Hanushek's most recent survey of the US literature was conducted, and to our knowledge, no published studies examining the wider international evidence have been undertaken over this period. It is therefore timely to re-evaluate more recent empirical contributions to gauge whether the wider body of recent international research has obtained more positive results, and if so, which inputs have been shown to have the greatest impacts on student achievement. There are two further motivations for examining contributions in the broader international literature. First, much of the survey work reviewed below concentrates exclusively on the US, whilst this review is primarily concerned with obtaining insights into the impact of education inputs on student outcomes in the UK. Given the considerably different institutional structures and input mixes between the two countries, examining a wider range of international studies provides a broader canvas from which to draw inferences. Second, and more importantly in light of the review's methodological focus, the examination of a wide range of high quality work can illuminate potential research strategies for future applied work in the UK. This point is particularly

¹⁹ Kirjavainen and Loikkanen find that there are significant changes in the magnitude of their estimates.

relevant, given the weak methodological basis underlying many existing UK studies (see Section 3).

This section is sub-divided into four major parts. The first part describes the two main methodological survey approaches used in evaluating input effects on student outcomes - vote counting and meta-analysis - and the results obtained by these survey techniques. The second section examines a set of 'high quality' contributions to the literature, post-dating Hanushek's most recent review. The 'high quality' label refers to papers that have been published in key journals. Some working papers are also included, but only those by world-renowned researchers that have made a serious attempt to overcome the methodological shortcomings highlighted in Section 1²⁰. The third part looks at the evaluation of specific policy interventions and the final part discusses conclusions.

2.1 Reviewing the reviews

The most well known surveys mapping education inputs to student achievement have been conducted by Hanushek (1986, and his frequent reappraisals, 1989; 1996; 1997a). Not surprisingly the stark oft-cited conclusion that "there is no strong systematic relationship between school expenditures and student performance"²¹ emphasised in Hanushek's studies has provided considerable controversy and has undoubtedly helped to fuel an extensive body of subsequent research.

Hanushek's conclusions were ostensibly based on what is described in the meta-analysis literature as vote counting: where the numbers of positive and negative statistically positive coefficients are summed and a distribution obtained. Re-evaluating Hanushek's original work, Hedges, Laine and Greenwald (HLG: 1994) called into question the validity of Hanushek's results and survey methodology²². The HLG critique is two pronged. First, they point out there are more positive than negative results in Hanushek's sample. Indeed, if the chance of being positive or negative were even, the odds of observing so many positive estimates would be less than one in a million.

The more substantive aspect of the critique uses meta-analysis techniques to formally evaluate the relationship between education inputs and student outcomes. HLG's results provide strong support for a robustly positive relation between student achievement and various inputs in the educational process. In particular, they found expenditure per pupil to be a robustly significant factor and that the mean coefficient was sufficiently large to be of practical importance²³. Of the other factors analysed, teacher experience was found to be the most consistently significant measured input. Furthermore, pupil-teacher ratios and teacher salaries gave mixed but generally significant results (although those results differed from grade to grade) while teachers' education was 'incorrectly' signed throughout.

Hedges *et al*'s criticism of Hanushek's findings has received recent affirmation by Dewey, Husted and Kenny (DHK: 2000). DHK use a more recent sample than that considered by either Hanushek (1986) or HLG (1994), including 33 papers and 127 regressions²⁴. DHK argue that the inclusion of income, because it is a demand-side factor,

²⁰ Relevant papers were obtained from three sources: the National Bureau of Economic Research Working Paper, the Princeton Industrial Relations Section Working Papers, and the IMF Working Paper series, although other working papers series were considered.

²¹ The initial statement being found in Hanushek (1986, p.1162).

²² A more recent update of this work is Greenwald, Hedges and Laine (1996). Their conclusion from this later meta-analysis was that "...school resources are systematically related to student achievement and that these relations are large enough to be educationally important." (Greenwald *et al*, 1996, p.385).

²³ Specifically, the coefficient value of 0.014 suggested that an increase in expenditure per pupil of US\$500 would be associated with a 0.7 standard deviation increase in a student's outcome.

²⁴ While the sample of papers is greater than those examined by Hanushek (1986) - DHK include 28 of the 33 papers examined by Hanushek (1986) and 18 additional papers - it is considerably smaller than that used in Hanushek's most recent review (Hanushek, 1997a) which includes 90 publications. The most recent paper

leads to inappropriate specifications, and that this is an important factor driving the lack of relationship in previous work²⁵. To test this hypothesis they sub-divide their sample of studies into 'good' and 'bad' estimates on the basis of two criteria. First, 'good' studies include a variable to capture parental input secondly they exclude parental income and other measures of socio-economic status which DHK argue are at the root of the mis-specification problem. Hence, those studies that exclude a parental input and include parental income and other socio-economic status measures are defined as 'bad' studies. Of the 127 studies in the complete sample, only about a quarter of the estimates are defined as 'good' under DHK's two selection criteria.

Table 2.1 provides a summary of the findings of Hanushek (1986 and 1997) and DHK (2000). DHK find that about 41% of the results of the 'good' studies are positive and significant, in contrast to the 15.4% and 19.7% in Hanushek (1986) and Hanushek (1997a) respectively. While the DHK study is informative, their comparison between 'good' and 'bad' studies does not universally suggest that 'good' studies are more likely to find a positive relationship between inputs and outcomes. In particular, expenditure per pupil was positively significant more often in 'bad' than in good studies, and the results from pupil teacher ratios are similar in both sub-samples.

A direct comparison between Hanushek's work and that of DHK is however complicated by two factors. First, DHK include variables not similarly classified by Hanushek, namely 'other teacher characteristics' and 'school size'. Second, the two samples differ substantively in the proportion of positive and significant results evaluated. This can be seen by comparing the proportion of positive and significant findings in the TOTAL row of Table 2.1. This shows that the Hanushek (1986, 1997) samples contain roughly a half and two-thirds respectively of the number of positive and significant results, as compared to the complete (ALL) DHK sample²⁶. To account for these differences the DHK analysis is recalculated by excluding 'other teacher characteristics' and 'school size', and comparing the results between the 'good' and 'bad' specification categories²⁷. The resulting recalculation suggests that there is practically no difference between the results of misspecified 'bad' studies and those of correctly specified 'good' studies, once factors excluded from Hanushek's analyses are similarly excluded from the DHK sample²⁸. In other words, including parental income does not seem to be the cause of the insignificant and mixed results that have been found in the literature.

included in the DHK analysis was published in 1996. It is unclear why the authors choose to link their study to Hanushek's earlier sample.

²⁵ See Section 1.1.1 and Mayston (1996) for an indepth treatment of the identification problem that occurs where educational inputs are endogenously determined by optimising behaviour on the part of schools where demand and supply factors are not separated.

²⁶ With the exception of pupil-teacher ratio, DHK's sample also has a lower proportion of negative and significant studies.

²⁷ The results are calculated using DHK Table 1. A more direct check of the robustness of their results would be to re-examine the Hanushek (1997a) sample using the DHK criteria.

²⁸ The resulting difference may be partially reconciled by Hanushek's finding that teacher test scores, which are included in DKH's 'Other teacher characteristics' were found to be positively significant in 37% of the studies he surveys (Hanushek, 1997a).

Table 2.1: Vote count analysis: comparison between Hanushek (1986, 1997) and Dewey, Husted, and Kenny (2000)

	<i>Percent with significantly positive coefficients^a</i>		Dewey, Husted and Kenny (2000)		
	Hanushek (1986)	Hanushek (1997)	All	Good	Bad
Expenditure per pupil	13	27	51.2	38.5	56.6
Teacher per pupil	9	15	25.8	29.4	24.4
Teacher education	6	9	28.1	25.0	28.8
Teacher experience	33	29	45.3	52.0	41.5
Teacher salary	9	20	45.4	00.0	45.4
Other teacher characteristics			43.4	52.6	40.4
School size			22.9	38.1	11.3
TOTAL	15.4	19.7	30.2	41.1	27.5
TOTAL (excluding 'school size' and 'other teachers characteristics') ^b			37.3	38.5	36.8

a Results are recalculated at the 5% significance level to facilitate comparison with Hanushek (1986 and 1997).

b Calculations are made using data contained in Dewey *et al*, Table 1.

There are a number of further areas of potential concern about the robustness of DHK's findings. First, DHK appear to take no account of 'aggregation bias' and, Section 1.1.2.a showed aggregation bias has been consistently identified as a factor spuriously raising the proportion of positively significant input coefficients²⁹. In common with other survey work, most of the methodological difficulties associated with the underlying studies examined are not considered by DHK³⁰. Yet vote counting analyses are generally quite sensitive to the studies that are included, and the modest number of 'good' relative to 'bad' studies in the DHK survey may exacerbate the seriousness of this problem. This may be illustrated by notionally adding an additional positively significant result to an input category. For example, adding a single positively significant result to 'teacher experience' results in an 8% rise in the proportion of positively significant results (from 25% to 33%), while if this exercise is repeated with the 'bad' sub-sample only a 2% increase result (from 17% to 19%)³¹.

The meta-analysis methodology applied in Dewey *et al* (2000), Hedges *et al* (1994) and Greenwald, Hedges, and Laine (1996) has not been without its critics, foremost among these being Hanushek. In their seminal work, Hedges *et al* (1994) attempt to placate their critics by assessing the potential weaknesses of their data set, acknowledging and mitigating obvious criticisms. It is instructive to review the points they raised, because (with the exception of the first) the criticisms apply to the underlying weaknesses of the studies surveyed and are therefore applicable to both the Hanushek and meta-analyses samples.

1. As Hanushek (1997, p.151) points out "combining test information is best motivated from taking a series of independent laboratories all providing results from a simple common experiment." However, the published estimates underlying surveys results are obtained

²⁹ Key references are contained in Section 1.

³⁰ Hanushek (1997a) does partially address 'quality' issues by examining a sub-set of results using value-added specifications.

³¹ It is not being suggested that the quality of papers underlying DHK research are better or worse in either the 'good' or 'bad' sub-samples, which would require a re-assessment of those papers underlying their work. Nor is it suggested that DHK's point that mixing demand and supply factors is analytically incorrect.

from studies which have used very different modelling strategies and comparison is problematic³².

2. It is widely held that “publication bias” reduces the quantity of published work that obtains either insignificant results, and/or perverse (‘wrongly’ signed) findings³³. Hedges *et al* (1994, p.12) argue that there is no reason to expect that publishers prefer ‘correctly’ signed coefficients, though it is “difficult or impossible to completely rule out selection effects”. While it is certainly true that proving the existence and direction of “publication bias” is a difficult task, finding an empirical article published in any journal that has no significant results is rare. Casual empiricism would also suggest that publication bias is more likely to bias upwards the proportion of positively significant results. Publication bias is a concern of both the meta-analysis and vote count approaches as both exclude unpublished material.
3. Given the extensive time period over which the papers included in most meta-analysis and vote-count studies are taken (generally over 30 years), it appears likely that the effect of various inputs on student achievement, and their relative importance, may have changed.
4. The final, and perhaps most compelling reason for scepticism, is that like all statistical analyses, the resulting findings are only as valid as the underlying data. Hence the statistical aggregation of work that is of a low methodological quality is likely to be uninformative³⁴.

Taken together, these factors suggest that the ‘positive’ results of some meta-studies can at best be taken as indicative, rather than conclusive. Rather than utilise either vote counting or meta-analysis, this review takes the more traditional approach of surveying a selection of individual studies, each of which has attempted to overcome some, but rarely all, of the methodological difficulties identified in Section 1. In doing so five of the most commonly utilised input variables are examined, namely: expenditure per pupil; class size or pupil-teacher ratio; teacher’s education, experience, and salaries³⁵. In addition, an assessment of the effects of a small selection of policy interventions is included.

2.2 Review of the recent international literature by input

In this sub-section recent (post-1997) contributions to the international literature, by input, are reviewed. For ease of reference two tables are presented for each input considered. The first table gives a basic summary of each paper, while the second highlights the methodological ‘soundness’ of each paper, against the criteria given in Section 1. Selected

³² In HGL’s defence they do attempt to address the problem by reassessing a sub-sample of results of independent studies. However, the results obtained differ substantively with those obtained for the full sample. Specifically, the coefficient values for per pupil expenditure fall by 29%, teacher experience by 57%, and the coefficients on the pupil-per-teacher ratio and teacher salaries change sign.

³³ In addition, publication bias may occur if researchers consider that publishers are less willing to publish insignificant or counter-intuitive results, and do not submit such work for publication.

³⁴ In particular, HLG explicitly point out that the studies incorporated in their sample are predominantly of a cross-sectional nature, and many of the papers included do not use value-added specifications. They also emphasize the weakness of many measures of socio-economic status in models in this field.

³⁵ In looking at this limited select of inputs it is acknowledged that other potentially important factors such as administrative inputs, school facilities, school size and leadership (Head Teacher characteristics for example) are not included. This is due to the lack of ‘high quality’ papers examining these issues. However where ‘high quality’ papers do include proxies for these factors, the results are reported within the key input sub-headings.

studies related to each input are then reviewed individually, and the sub-section concludes with a brief summary.

2.2.1 Expenditure per pupil

Table 2.2 summarises the results of four recent studies that examine expenditure per pupil as an independent variable. None of these studies examine expenditure per pupil at a disaggregated level and hence all are subject to aggregation bias concerns. Indeed, the apparent lack of attention in recent literature to the expenditure per pupil variable suggests that researchers and publishers have acknowledged the difficulty in interpreting the results of such studies where the expenditure variable is collected at A-level of aggregation above the pupil level (aggregation bias)³⁶.

Of the papers examined perhaps the most convincing is that of Figlio (1997a). Figlio's principal interest is in determining if caps on revenue and expenditure in certain states constrained educational resource allocation in these states, and whether this 'random' variation in educational resourcing can explain student achievement. His results indicate that expenditure reductions were harmful to students' achievement in mathematics, reading, science, and social studies. Furthermore, the impact of reduced expenditure was quantitatively important, with a revenue and expenditure cap having an impact on student achievement equivalent to a reduction in family income of US\$28,000. However, in addition to the 'aggregation criticism' (although Figlio's data is aggregated at the district rather than the state level which reduces the degree of bias) the paper does not provide a careful check on the robustness of the impact of reduced expenditure on outcomes, making it difficult to comment on the rigor of the results³⁷.

Dewey *et al* (2000) use instrumental variable techniques to infer a causal relationship between expenditure per pupil and SAT scores, and obtain positive results. However, the data they examine is highly aggregated (state level), and the sample size is relatively small (222 observations).

Gupta, Verhoeven and Tiongson (1999) examine the determinants of enrolment rates in a cross-country framework using instrumental variables. They found that countries that invest a greater proportion of national income in education have higher enrolment rates. Apart from the usual 'aggregation' criticism, there are a number of additional limitations associated with cross-country studies, such as measurement error, and omitted variables bias. It is also unclear if the Gupta *et al* (1999) results translate into an OECD country context, since the authors do not separate their sample into OECD and non-OECD countries.

Marlow (2000) focuses mainly on the effect of competition on school performance, where competition is measured by an index based on the number and concentration of different school districts within a particular county. His premise is that a larger number of school districts, with more equal shares of the market, will stimulate greater competition in that county and raise performance. This paper is unusual in that, although it suffers from aggregation bias, it does address the question of the endogeneity of educational expenditure by estimating a two-equation model. Marlow first models the determinants of per capita educational expenditure at primary and secondary school level. He then models the effect of this expenditure on Grade 4, 8 and 10 reading, writing and mathematics test outcomes. He finds no evidence that higher expenditure leads to better outcomes, and in fact many of his results suggest a statistically significant negative relationship between expenditure and outcomes. He also found that educational spending per capita was higher in counties with the

³⁶ For example, expenditure per pupil was the most common variable analysed in Hanushek (1986), and the third most common in Hanushek (1997a).

³⁷ The first part of the paper, which examines the impact of caps on revenue and expenditure on educational resourcing, is subject to careful robustness checks. Unfortunately, in the absence of similar checks concerning the proposed link between expenditure and student outcomes, it is difficult to interpret the results.

greatest monopoly power (fewest school districts). This latter result is consistent with the argument that higher expenditure does not lead to better outcomes because higher expenditure tends to flow to school districts, administrators, teachers and staff for reasons unrelated to performance.

Table 2.2: INPUT – Expenditure

Author	Statistical Technique	Location	Magnitude of key results – effect of 1% inc. in expenditure on outcome	Controls
Dewey et al. (2000)	OLS, IV	US	OLS: 0.15% IV: 0.12%	Standard controls (but excluding income)
Figlio (1997)	OLS, IV, Difference-in-difference	US	2.5 – 6.4% (at 5% and 10% significance levels)	Student, family, school controls
Gupta et al. (1999)	OLS, 2SLS (output is enrollment rates)	Cross-country	Primary+Secondary/ Total Educational Expenditure 0.2% ^a Educ. Exp./GDP 3.3% ^a	Population, urbanisation, child mortality
Marlow (2000)	SUR – Seemingly Unrelated Regression	County level data	Primary+Secondary Educational Expenditure Ranged from -0.01 to +0.002	Expenditure equation – per capita income, pop density, student share of pop, state and federal share of funding, race, public sector competition index Outcome equation – as above plus median education and expenditure

Notes:

1/ Results given are statistically significant at the 1% level or better (unless otherwise stated)

2/ Refer to student level data (unless otherwise stated)

3/ Where results are not significant co-efficient values are not reported.

4/ All outputs are test scores unless otherwise stated

^a Only 2SLS estimates are reported.

Overall, sufficiently significant concerns about the data and methodological tools employed in the four papers examined suggest that the ‘Hanushek view’ - that increased expenditure in itself does not raise student achievement - cannot be seriously challenged by the results of these studies.

Table 2.3: INPUT – Expenditure

Author	Methodological issue			
	1.1.1 Endogeneity	1.1.2a Aggregation	1.1.2b Functional form	Omitted variable bias
Dewey et al. (2000)	4	×	×	×/4
Figlio (1997)	×	×	×	×/4
Gupta et al. (1999)	4	×	4	×/4
Marlow (2000)	4	×	×	×/4

Key: 4 Attempts to overcome methodological difficulty; × No attempt to overcome methodological difficulty; × /4 Some attempt to overcome methodological difficulty.

2.2.2 Class size

The issue of class size remains a hotly contested one, by the public, and by professional educators. Reviews of this literature have suggested that smaller class sizes do not systematically lead to improved student achievement (Hanushek, 1997a). However, some recent and rigorous research findings seem to contradict this. Of the three most convincing studies considered here, that do find a positive link between smaller classes and student achievement, two are from the US (Krueger, 1999; Hanushek, Kain, and Rivkin, 1998), and the third from Israel (Angrist and Lavy, 1999). These positive findings are tempered by four papers that obtained negative (or insignificant) results (Hoxby, 1998; Cooper and Cohn, 1997; Goldhaber and Brewer, 1997; and Goldhaber, Brewer, and Anderson, 1999). Finally, Wright, Horn and Sanders (1997) find that smaller class sizes in themselves do not lead to higher achievement, but that the interaction effects are important. The relative merits of each of these studies is examined in turn.

Table 2.4: INPUT – Class Size

Author	Statistical technique	Location	Magnitude of key results – effect of 1% dec. in class size on outcome	Controls
Angrist and Lavy (1999)	Natural Experiment	Israel	5 th grade: 3.6% 4 th grade: 1.7-1.9%	Family background
Barro and Lee (1996)	SUR (Pupil-teacher Ratio)	Cross-country	2.7%	Cross-country controls
Cooper and Cohn (1997)	Stochastic Frontier/ OLS	US (S.Carolina)	1-4%	Background characteristics
Goldhaber et al. (1999)	OLS Panel (Pupil-teacher ratio)	US	-6.6% to -7.2%	Class, teacher, and school
Hanushek et al. (1998)	Panel	US (Texas)	0.001-0.05%	Fixed effects and background characteristics
Hoxby (1999)	Natural experiment	US (Connecticut)	Insignificant	District level fixed effects, time trends, cohort fixed effects, and background characteristics
Kirjavainen and Loikkanen (1998)	DEA (Pupil-teacher ratio)	Finland	25% of inefficiency explained	Background characteristics
Krueger (1999)	Experiment (Panel)	US (Tennessee)	7-9% level effect 1% growth p.a.	Background characteristics, school dummies
Krueger and Whitmore (1999)	Experiment (Panel)	US (Tennessee)	20% increased probability of taking college entrance test	Fixed effects and background characteristics

Notes:

1/ All outputs are test scores (unless otherwise stated).

2/ Results given are (i) statistically significant at the 1% level or better, and (ii) refer to student level data (unless otherwise stated).

3/ Where results are not significant co-efficient values are not reported.

^a Range of co-efficient values reported where they were significant at the 5% level (4 of the 12 estimations).

The Krueger (1999) paper analyses the results of a unique random experiment undertaken in Tennessee between 1985/86 and 1988/89. The central benefit of experimental data is that, despite some limitations, it by and large ensures that the class size variable is not endogenously determined (Section 1.1.1d).³⁸ In the Tennessee experiment, children from kindergarten and into grades 1 to 3 were randomly allocated to large (22-24 pupils) and small (14-16 pupils) classes. Teachers were also assigned to classes on a random basis. Students' progress was assessed using a standardised test and, after the first year, children in smaller classes did significantly better than their peers in larger classes (by about 5-8 percentile points). The benefits were greater for minority and poorer students. Krueger found that the most substantial gains occurred in the initial year of class size reduction, with the differential between children in small and large classes increasing by a single percentage point in

³⁸ Krueger checks for this finds no evidence of "Hawthorne effects", *i.e.* systematic changes in teacher behaviour simply as a result of being in the experiment.

subsequent years. A limitation of the work of Krueger (1999) is that, by not tracking students in later years, he is unable to gauge whether the gains of smaller class sizes ‘fade out’ over time.

A promising subsequent analysis by Krueger and Whitmore (1999) suggests that the student achievement associated with the STAR experiment had permanent effects. Krueger and Whitmore test this by examining whether the probability of taking college entrance exams was higher for children who were previously in small classes. The evidence they provide indicates that those who were in smaller class sizes were 20% more likely to sit the Standard Aptitude Test or the Texas Assessment of Academic Skills tests. Promising though these results may be, the authors acknowledge that they are based on an incomplete sample and should be viewed as preliminary.

Hanushek, Kain, and Rivkin (1998) provide alternative estimates of class size effects using the Harvard/UTD Texas Schools Project database. There are two key advantages associated with this database. First, it is considerably larger than any previous data set used in educational production function estimation (including 3,000 schools with over 200,000 students per cohort for 4th, 5th and 6th grades), allowing precise coefficient estimation. This student level dataset is also linked to a variety of demographic and student background measures. Second, the repeated nature of the study, which has been running since 1993, permits the estimation of value-added fixed effects models that explicitly attempt to overcome the problems of endogeneity and omitted variable bias. Hanushek *et al* (1998) find that class size effects are statistically significant for 4th and 5th formers but not for 6th formers. However, although significant, the magnitude of the effects are considerably smaller than those obtained by Krueger (1999), and account for less than 0.1% of the total variation in student achievement.³⁹ However, Hanushek *et al*'s results may well underestimate the total impact of class size on student outcomes, given that the STAR research suggests a one off level increase in pupil achievement in the initial year of the experiment, and such level gains are not captured by value-added specifications.

The third study that finds robust gains in student achievement from smaller classes, is an examination of class size reductions in Israel in the early 1990s by Angrist and Lavy (1999). As discussed in Section 1.1.1c, Angrist and Lavy cleverly use Maimonides' rule - that class sizes cannot exceed 40 - which was enshrined in Israeli educational policy over the period examined - to identify an exogenous or random discontinuity that is used to instrument for class size changes. As ever, there are a number of technical limitations to Angrist and Lavy's work⁴⁰. However, the prime issue is that the nature of the instrumental variable used means that Angrist and Lavy primarily consider the effect of class size changes around the maximum limit of 40. Extrapolating these results to class size changes in OECD countries, which generally have much lower average class sizes, is problematic.

The problem of comparing estimates from OECD and non-OECD countries also pertains to a cross-country study undertaken by Barro and Lee (1996). This examined a sample of 58 countries between 1967 and 1991 and found that class size, proxied by the pupil-teacher ratio, had a significant impact on student outcomes in international tests. However, data limitations do not allow Barro and Lee to overcome the majority of potential methodological difficulties identified in Section 1, as is shown in Table 2.5.

Hoxby (1998) uses two quasi-experimental techniques in a panel framework to examine the influence of class size on test scores in Connecticut district schools. She finds there to be no significant impact of class size on student achievement. Hoxby's work is noteworthy since, it is extremely methodologically rigorous, and is the only study obtaining insignificant results which focuses exclusively on the class size issue. In addition, Hoxby provides a convincing critique of the Angrist and Lavy (1999) paper.

³⁹ Krueger uses statistical techniques to adjust for non-random attrition.

⁴⁰ See Hoxby (1998) for a detailed discussion.

Examining the National Educational Longitudinal Study (NLES) of 1988, Goldhaber and Brewer (1997) found that class size was significant but had the ‘wrong sign’, *i.e.* suggesting that larger classes are associated with better student outcomes. In a subsequent paper using the same data, Goldhaber, Brewer, and Anderson (1999) found a (not surprisingly) similar result. On the other hand, Cooper and Cohn (1997), using stochastic frontier estimation, found that smaller class sizes reduce student test scores. The resulting class size estimates are less robust than those obtained in Goldhaber and Brewer (1997), or Goldhaber *et al* (1999), being significant in only 8 of the 12 specifications examined at the 5% significance level. Perhaps the major limitation of all these studies is the lack of attention paid to the endogeneity issue, a problem shared by much work in this area and which Goldhaber and Brewer (1997, p.513) explicitly acknowledge. Goldhaber *et al* also highlights the low explanatory power of observable student characteristics, finding them to account for a mere one percentage point of total variation. While the result is disappointing, a clear limitation of this study is the lack of any underlying structural model to determine the inter-relationships between the large number of explanatory variables examined. The work of Goldhaber and Brewer, Goldhaber *et al*, and Cooper and Cohn all highlight the need for using theory to motivate empirical modeling and testing.

Finally, while not explicitly examining a structural model of schooling, a paper by Wright, Horn and Sanders (1997) does suggest that, although class size in itself may not be important, the interaction of class size with other input factors does have a significant effect on student outcomes (Section 1.1.1b). Using the Tennessee Value-Added Assessment System database in two regional subsamples, and applying panel techniques, the authors observe that class size *per se* does not lead to higher achievement. However, allowing for student heterogeneity (proxied by sub-dividing students into three achievement sub-groups), and including interactions (such as the interaction of student prior achievement with class size) produces highly significant effects - in both fixed effects and random effects models. The findings are particularly interesting in that they provide a perhaps more intuitively appealing method to analyse class sizes, by operationalising notions commonly stated but rarely explicitly tested in the literature. In particular, there appears to have been little or no debate over whether smaller class sizes in themselves provide the *opportunity* to improve achievement, or whether the effect of class size on the opportunity to learn is largely determined by the actions of teachers and other factors exogenous to the classroom⁴¹. The finding that the nature of the students, school system, and most importantly the class teacher, all impact on the way smaller classes generate better outcomes is an encouraging one⁴².

In summary, while the higher quality evidence, based on more satisfactory methodological strategies, is supportive of reduced class sizes being a positive influence on student outcomes, the magnitudes of the effects found do not appear large enough to justify increased expenditure. Krueger’s analysis suggests that a 1% decrease in class size would lead to A-level gain of 4% followed by a 1% growth rate per annum, which appears to be an upper bound on the potential return from reducing class size. In the context of the Tennessee experiment, Krueger undertakes a relatively simplistic cost-benefit analysis and shows that the costs of class size reduction are roughly equal to the benefits. By implication, the smaller

⁴¹ See Hanushek (1999, p.117) for a clear articulation of the variety of sources through which class size affects student outcomes.

⁴² While the Wright, Horn and Sander method is a promising approach, there are a number of difficulties associated with this paper. First, the notion of ‘school system’ is undefined. Second, while the authors find that heterogeneity and class size (via interacting the two variables) lacks explanatory power the mechanism through which teachers interact with schooling, student heterogeneity and class size is unclear. This reflects the omission of other combinations of interaction terms, but more seriously the lack of an analytical framework detailing the mechanisms that drive the results. Third and somewhat unconventionally, the authors only report the significance levels and not the value of the respective coefficients. So while it appears that the interaction terms are important, the magnitude of their importance is unclear. Finally, the authors provide no indication as to the statistical validity of the models examined through specification testing.

effects found in other studies suggest that class size reductions in themselves are not a cost-effective means of enhancing student outcomes.

Table 2.5: INPUT - Class Size

Author	Methodological issue			
	<i>1.1.1 Endogeneity</i>	<i>1.1.2a Aggregation Bias</i>	<i>1.1.2b Functional form</i>	<i>Omitted variable bias</i>
Angrist and Lavy (1999)	4	4	4	4
Barro and Lee (1996)	✗	✗	✗	✗/4
Cooper and Cohn (1997)	✗	4	✗	✗/4
Goldhaber and Brewer (1997)	✗	4	✗	4
Goldhaber et al. (1999)	✗	4	✗	4
Hanushek et al. (1998)	4	4	✗/4	4
Hoxby (1999)	4	4	4	4
Kirjavainen and Loikkanen (1998)	✗	✗	4	✗/4
Krueger (1999)	4	4	4	4
Krueger & Whitmore (1999)	4	4	4	4
Wright, Horn and Sanders (1997)	✗	4	✗	4

Key: 4 Attempt to overcome methodological difficulty; ✗ No attempt to overcome methodological difficulty; ✗ /4 Some attempt to overcome methodological difficulty.

2.2.3 Teacher characteristics

Tuition is a highly labour intensive and expensive process. Indeed, Audit Office estimates show that teachers' salaries account for 70% of total school education expenditure in the UK⁴³. Perhaps the most disturbing aspect of the literature is therefore the finding that measurable teacher characteristics appear to have little bearing on achievement. The Hanushek (1997a) survey gives the percentage of studies finding significant results for three of the most commonly included teacher characteristics - teacher's education (9% positive and significant), experience (29%), and salary (27%). Consequently the results are far from being consistently positive and significant. In addition, the least satisfactory performer, teacher education, had a negative impact in 5% of the studies surveyed. However, a re-examination of the evidence, using meta-analysis techniques, suggested a positive relationship between some teacher characteristics (experience) and outcomes (Greenwald *et al*, 1996).

Similarly recent research by Hanushek, Kain and Rivkin (1998), decomposing the determinants of student achievement, found not only that schools have a potent impact on achievement differences, but that teachers appear to be the most important specific factor⁴⁴. We examine each of the three most popularly employed measures of teacher quality, to ascertain whether other studies support Hanushek *et al*'s appraisal.

⁴³ *Op cit*, Mayston and Jesson (1999), pp.17.

⁴⁴ Specifically, Hanushek *et al* find a lower bound of 7.5% of the total variation in student achievement is due to teachers.

2.2.3a Teacher experience

Hanushek *et al* (1998) found significant positive effects on achievement for teachers with up to 2 years experience, as compared to teachers with no experience (with the exception of 4th and 5th form mathematics). However, they found no effects from greater levels of teacher experience. The magnitude of the results are considerably greater than the effects of the other input measures they considered, and were about 20 times larger than class size effects. In contrast, Krueger (1999) found gains of less than half those obtained by Hanushek *et al* (1998), when comparing teachers with no experience with experienced teachers. Furthermore, using a quadratic model Krueger finds that the positive impact of having greater experience peaked after 20 years.

Table 2.6: INPUT – Teacher Experience

Author	Statistical technique	Location	Magnitude of key results – effect of 1% inc. in teacher experience	Controls
Cooper and Cohn (1997)	Stochastic Frontier estimation	US (S.Carolina)	Insignificant	Background characteristics
Dewey et al. (2000)	OLS, IV	US	OLS: -0.008% to 0.04% IV: 0.08%	Background characteristics
Hanushek et al. (1998)	Value-Added (Panel)	US (Texas)	-7% to -15% ^a -4% to -10% ^b	Class, teacher, and school controls
Krueger (1999)	Experiment (Panel)	US (Tennessee)	3%	Background variables, school dummies

Notes:

1/ All outputs are test scores (unless otherwise stated).

2/ Results given are (i) statistically significant at the 1% level or better, and (ii) refer to student level data (unless otherwise stated).

3/ Where results are not significant co-efficient values are not reported.

^a Maths: Measures negative effect from % of teachers with just one or zero years experience respectively.

^b Reading: Measures negative effect from % of teachers with one or zero years experience respectively.

As summarised in Table 2.6, other studies surveyed in this paper (other than Dewey *et al*) all found that teacher experience lacked explanatory power. However, these alternative studies do not directly contradict Hanushek *et al* since the teacher experience proxy used in each of these studies was *total* teacher experience, and so unlike Hanushek *et al* they do not differentiate between experience obtained early in a teacher's career and a teacher's total teaching experience.

Table 2.7: INPUT - Teacher Experience

Author	Methodological issue			
	<i>1.1.1 Endogeneity</i>	<i>1.1.2a Aggregation</i>	<i>1.1.2b Functional form</i>	<i>Omitted variable bias</i>
Cooper and Cohn (1997)	✗	4	✗	✗/4
Dewey et al. (2000)	T	Y	Y	YU
Goldhaber and Brewer (1997)	✗	4	✗	4
Hanushek et al. (1998)	4	4	✗	4
Krueger (1999)	4	4	4	4

Key: 4 Attempt to overcome methodological difficulty; ✗ No attempt to overcome methodological difficulty; ✗ /4 Some attempt to overcome methodological difficulty.

2.2.3b Teacher education

As far as teacher education is concerned only three of the studies considered find significant results. In particular Goldhaber and Brewer (1997) detect robust results that being taught by a teacher with a degree in mathematics has a positive impact on pupils' mathematics scores. By contrast however, Hanushek *et al* (1998) found that 4th form students appear to suffer through having more highly educated teachers. Neither study found that, outside mathematics, there is any discernable relation between teacher education and outcomes - a result that is consistent with Goldhaber *et al*'s recent re-examination of the NLES database. The only paper that appears to have had some success with an 'aggregate' teacher education variable is Cooper and Cohn (1997), however their results are not robust to all the specifications.

Table 2.8: INPUT – Teacher Qualifications

Author	Statistical technique	Location	Magnitude of key results	Controls
Cooper and Cohn (1997)	Stochastic Frontier estimation	US	0.5-2.5% ^a - effect from Masters degree	Background characteristics
Dewey et al. (2000)		US	Insignificant	Background characteristics
Goldhaber and Brewer (1997)	OLS Panel	US	2.3% - effect from Certificate in Math ^b 0.82% - effect from BA Degree in Math ^b	Class, teacher, and school controls
Goldhaber et al. (1999)	OLS Panel	US	Insignificant	Class, teacher, and school controls
Hanushek et al. (1998)	Value-Added (Panel)	US (Texas)	-4% - effect from Masters degree	Class, teacher, and school controls
Krueger (1999)	Experiment (Panel)	US (Tennessee)	Insignificant	Background variables, school dummies

Notes:

1/ All outputs are test scores (unless otherwise stated).

2/ Results given are (i) statistically significant at the 1% level or better, and (ii) refer to student level data (unless otherwise stated).

3/ Where results are not significant co-efficient values are not reported.

^a 7 of 12 OLS estimations significant at the 5% level, none are significant in the frontier estimations.

^b Results from Random effects model reported.

Table 2.9: INPUT - Teacher Qualifications

Author	Methodological issue			
	1.1.1 Endogeneity	1.1.2a Aggregation Bias	1.1.2b Functional form	Omitted variable bias
Cooper and Cohn (1997)	✗	4	✗	✗/4
Dewey et al. (2000)	4	✗	✗	✗/4
Goldhaber and Brewer (1997)	4	4	✗	4
Goldhaber et al. (1999)	✗	4	✗	4
Hanushek et al. (1998)	4	4	✗	4
Krueger (1999)	4	4	4	4

Key: 4 Attempt to overcome methodological difficulty; ✗ No attempt to overcome methodological difficulty; ✗ /4 Some attempt to overcome methodological difficulty.

One additional study that is worth mentioning, even though it formally falls outside the remit of this review since it was published in 1994, is Monk (1994). Monk found evidence that teacher subject preparation, *i.e.* how many courses a teacher took in the subject area being taught, is positively related to students' performance in some subjects, namely mathematics and science. He also found evidence that teachers who had taken course work in pedagogy

had a positive impact on their students' performance. In general he concluded that teacher preparation does make a difference. The study is noteworthy because it uses very detailed data, explores possible non-linear relationships between the input and output variables, investigates threshold effects and takes into account possible interactions. However, this study does not contradict the more recent work reviewed above, since it too found little evidence that more aggregate measures of teacher characteristics (degree levels, college credits or years of teacher experience) have a systematically positive impact on pupil performance.

2.2.3c Teacher salaries

A key recent paper examining teacher salaries, Hanushek, Kain and Rivkin (1999), found that when student fixed effects and teacher mobility is accounted for, higher salaries exert a positive influence on student achievement. But implausibly, the evidence suggests that higher salaries have the strongest effects on tenured experienced teachers, and not on young untenured teachers⁴⁵. Dewey *et al* obtain positive and robust results that relative salary differentials matter in determining student outcomes. Other analyses, summarised in Table 10, find no significant relationship.

Table 2.10: INPUT – Teacher Salaries

Author	Statistical technique	Location	Magnitude of key results – the effect of a 1% inc. in teachers' salaries	Controls
Barro and Lee (1996)	SUR	Cross Country	Insignificant	Cross-country controls
Dewey et al. (2000)	OLS, IV (seemingly unrelated regression)	US	OLS:0.04% ^a IV: 0.07%	Background characteristics
Goldhaber et al. (1999)	OLS Panel	US	Insignificant	Class, teacher, and school
Hanushek et al. (1999)	Value-Added (Panel)	US (Texas)	0.76-1.2%	Class, teacher, and school fixed effects, and standard controls

Notes:

1/ All outputs are test scores (unless otherwise stated).

2/ Results given are (i) statistically significant at the 1% level or better, and (ii) refer to student level data (unless otherwise stated).

3/ Where results are not significant co-efficient values are not reported.

^aIn one of three specifications reported.

⁴⁵ Figlio (1997b), using probit analysis, complements Hanushek *et al*, finding that teacher salaries lead to the recruitment of higher quality teachers (which Figlio measures using undergraduate college selectivity and subject matter expertise) within and between local labour markets. Figlio's results are of a higher magnitude than Hanushek *et al*, but he does not examine whether or not schools with higher quality teachers obtain better outcomes.

Table 2.11: INPUT - Teacher Salaries

Author	Methodological issue			
	1.1.1 Endogeneity	1.1.2a Aggregation bias	1.1.2b Functional form	Omitted variable bias
Barro and Lee (1996)	✗	✗	✗	✗/4
Dewey et al. (2000)	4	✗	✗	✗/4
Goldhaber et al. (1999)	✗	4	✗	4
Hanushek et al. (1999)	4	4	✗	4

Key: 4 Attempt to overcome methodological difficulty; ✗ No attempt to overcome methodological difficulty; ✗ /4 Some attempt to overcome methodological difficulty.

To sum up, there is some robust evidence that teacher experience and teachers' salaries have significant effects but that teachers' education levels (with the exception of teachers with qualifications in mathematics) do not. The results also provide considerable support for the use of more refined measures of these characteristics, such as specifically examining initial years of teacher experience, and looking at teacher qualifications by subject area.

While education, experience and salaries are the teacher characteristics most commonly examined in empirical work, reflecting to a large degree the availability of data, few educators would argue that these are the only relevant factors. With the notable exception being Goldhaber and Brewer (1997) there appears to have been little work done on examining alternative factors. Their paper is important because it evaluated the effect of both observed and unobserved teacher and school characteristics on students' 10th grade mathematics scores. They found that teacher behaviours and techniques may be more important than simple resource measures. Specifically, teachers who felt well prepared, who had control over lesson content, who spent less time maintaining order, and who used oral questions frequently and emphasised problem solving had a positive effect on pupils. The causality and endogeneity of some of these behavioural variables is questionable, but these results do suggest that researchers might need to focus more on qualitative aspects of teachers and schooling inputs.

2.3 Policy interventions

While outside the scope of most previous surveys, the effect of specific policy interventions on educational achievement is clearly of great interest to policy makers. Four such studies that attempt to overcome methodological and conceptual shortcomings, and hence meet the reviews 'high quality' criteria are included in this Section. The four are centered on the returns to a number of varied policy interventions in the US: Head Start, Teacher Salary Incentive Schemes in South Carolina, a similar personal incentive scheme in Dallas, and an education voucher scheme (the Milwaukee Parental Choice Program).

2.3.1 Head start

The Head Start program was first implemented in the 1960s and is the most widespread US pre-school intervention. While there a number of papers that have examined the returns to Head Start, the seminal paper in the area comes from Currie and Thomas (1995). They used comparisons between siblings in the same household (fixed effects models) to evaluate the educational effects of attending a Head Start pre-school programme⁴⁶. The fixed effects

⁴⁶ There are a considerable number of papers examining the Head Start Program. The majority of the more careful contributions are contained and summarised in Currie and Thomas (1995). Only Currie and Thomas

approach was designed to overcome the bias caused by unobservable factors influencing sample selection, since Head Start is not a random intervention. In particular, children who attend a Head Start programme may not be representative if there are unobservable factors (such as motivation and ambition) which make some parents more likely to enrol their children in the program. Briefly, they found a positive effect on test scores from Head Start for white pupils but no effect for African-American students. Further investigation suggested that both groups of pupils benefited equally in the early school years from being in Head Start, but this positive effect had disappeared for African-American students by the age of 10. The authors found some evidence that this could be due to either lower quality Head Start programmes for African-Americans, or the disproportionately low quality of schooling subsequently experienced by this minority group. In a recent paper, Currie and Thomas (1995) provide evidence that lower *post-intervention* school quality, in the sense that African-American students attended schools with worse test scores, was indeed a factor.

2.3.2 Teacher salary incentive schemes: South Carolina

Using OLS and stochastic frontier techniques, Cooper and Cohn found that two salary incentive programs, designed to reward teachers and enacted in South Carolina, had a positive influence on student achievement. The first plan (Plan 1) identified teacher performance on the basis of four criteria, namely: teacher attendance, performance evaluation according to state criteria, self-improvement through attendance in at least one 'self-improvement activity' and their students' achievement. All teachers selected for an award received no less than US\$2,000 but no more than US\$3,000. Under the second plan (Plan 2) school districts allocated one-third of the teacher bonus to individual teachers and two-thirds to the school. The intent of Plan 2 was to reward high performance schools and encourage 'team work' among teachers. While there are a number of methodological shortcomings to the study (see Table 2.13 below) the resulting school-level estimates suggest that both plans significantly enhanced student outcomes, but that the purely individual based plan (Plan 1) led to relatively greater student achievement. The results were robust in both stochastic frontier and OLS estimations.

Table 2.12: INPUT- Policy Interventions

Policy Intervention	Author	Statistical technique	Magnitude of key results	Controls
Head Start	Currie and Thomas (1995)	Panel	7% (initial gains)	Household fixed effects and background characteristics
Teacher Pay Incentives (S.Carolina)	Cooper and Cohn (1997)	OLS Stochastic Frontier estimation	Plan 1: 1.6-3.4% Plan 2: 0.8-2.9%	Background characteristics
Dallas Incentive Scheme	Ladd (1999)	Panel	1.0 -1.7% enhanced pass rate for TAAS ^a	Time and city dummies; individual characteristics
Voucher scheme	Rouse (1998)	Panel	1.5-2.3%	Individual fixed effects

Notes:

1/ All outputs are test scores (unless otherwise stated).

2/ Results given are (i) statistically significant at the 1% level or better, and (ii) refer to student level data (unless otherwise stated).

3/ Where results are not significant co-efficient values are not reported.

^a Texas Assessment of Academic School test.

(1995, 1999) is considered here since their work is notable for its careful methodological treatment of the non-randomised nature of the Program.

2.3.3 Accountability and incentive scheme: Dallas

Ladd (1999) explored the impacts of the Dallas program implemented in 1991/92. Under that program personnel at the most effective schools received financial rewards⁴⁷. The scheme is somewhat similar to Plan 2 implemented in South Carolina, in that it rewarded teachers in successful schools, but was more wide ranging by rewarding other staff members too. Specifically, principals and teachers typically received US\$1,000, while other staff (such as secretaries, and cleaners) received about US\$500, and an additional US\$2,000 was allocated to the school itself.

Using panel techniques, via the inclusion of city, school and time dummies, interactions of these dummies for both Dallas and four neighboring Texan cities (Fort Worth, Houston, Ysleta, and El Paso), and controls designed to capture student characteristics, Ladd's central finding was that the pass rate of 7th form students was significantly improved. Overall the gains were in the order of 10 to 20% of the state average. In addition, Ladd found that the reforms lowered drop out rates, and that Head Teacher turnover increased dramatically (from 2.4 to 24.6%) in least effective schools⁴⁸. Ladd tests whether an overall increase in the level of educational resources available to Dallas schools caused this result, and whether another (wider) state program targeted at low performing schools could have affected his results. He found neither factor to be important. A puzzling aspect of Ladd's findings is that pass rate increases were greatest in the initial year of the study, which implies that the program was only partially successful, since incremental gains in subsequent years would be expected in a truly successful intervention.

2.3.4 Voucher systems: Milwaukee Parental Choice Program

Rouse (1998) examines the Wisconsin voucher system that was initially implemented in 1990⁴⁹. Using unsuccessful applicants⁵⁰ and children enrolled in Milwaukee public schools as two comparison groups, Rouse undertook two separate analyses, applying panel and instrumental variable techniques⁵¹. Rouse consistently observed that mathematics marks were enhanced for students using the voucher scheme to attend private schools (by 1.5-2.3%), while reading was largely unaffected⁵². However, in a more recent working paper, Rouse (1999), alters the interpretation of the earlier results. Rouse examines a different comparison group in the form of different types of public schools. The comparison suggests that certain higher quality state schools ("P-5" schools), which have greater funding and lower class sizes than normal state schools, tended to perform as well as Choice Schools. This suggests that a

⁴⁷ Effectiveness being determined on the basis of two tests: the Texas Assessment of Academic Skills, and the Iowa Test of Basic Skills. 20% of the Dallas schools received financial rewards.

⁴⁸ Due to a relative lack of turnover in poor performing schools Ladd interprets this as a positive factor.

⁴⁹ In a recent study Peterson, Myers, Howell and Mayer (1999) found that a voucher system applied in New York enhanced student achievement. This work is not considered here in detail since there is evidence that the experiment was non-random and the authors do not adequately account for this via statistical correction techniques, it should be noted though that their results are consistent with Rouse (1998).

⁵⁰ Those parents who were rejected from the scheme.

⁵¹ There are two other studies of the Milwaukee Parental Choice program. The first concludes that there were no relative gains for choice students (Witte, Thorn and Sterr, 1995 who compare choice students with a random sample of Milwaukee public school students). The second, by Du, Green, and Peterson (1997), finds that students made statistically significant gains by their third and fourth years in the program in both reading and maths. Rouse (1998) is superior, including both these approaches but also using greater controls in the form of fixed effects and IV models.

⁵² Rouse (1998) rightly stresses that limitations in the underlying data, and the small-scale nature of the Program, make inferences on the effects of a more widespread programme problematic. Furthermore, Goldhaber, Brewer, Eide, and Rees (1999) argue that Rouse does not explicitly account for the possibility of non-random attrition.

voucher system is only one means to achieve gains in student achievement and provides indirect evidence that reduced class sizes enhance student outcomes⁵³.

To conclude; the four policy interventions which have been reviewed are found to robustly impact on pupil attainment.

Table 2.13: INPUT- Policy Interventions

Author	Methodological issue			
	<i>1.1.1 Endogeneity</i>	<i>1.1.2a Aggregation bias</i>	<i>1.1.2b Functional form</i>	<i>Omitted variable bias</i>
Cooper and Cohn (1997)	✗	4	✗	✗/4
Currie and Thomas (1995)	4	4	✗	4
Ladd (1999)	✗	4	4	✗/4
Rouse (1998)	4	4	✗	4

Key: 4 Attempt to overcome methodological difficulty; ✗ No attempt to overcome methodological difficulty; ✗ /4 Some attempt to overcome methodological difficulty.

2.4 Concluding comments

Taken as a whole, the ‘high quality’ international research findings suggest that some measurable school inputs do matter. These include class size, teacher experience and teacher salaries. However, the magnitude of the effects found has been quite small. The evidence on specific educational interventions is more optimistic, as most schemes considered generated substantially improved student performance.

Hence the literature as it stands does not imply that simply increasing expenditure on education would be appropriate in all cases. Recent research, especially Hanushek, Kain, and Rivkin (1998 and 1999), reinforces the message that examining the *magnitude* of any school resource effects found is essential. These results are more positive than Hanushek’s (1997) original assessment that school-resourcing levels may not have any impact on student performance. Perhaps the key reason underlying the subtle change in the ‘Hanushek view’, is that better data have become available, which has allowed more recent studies to use more credible empirical techniques. This underlines the need for good quality methodological work in this field. It would be overly optimistic to argue that a limited series of even well designed empirical studies could lead to definitive conclusions; however, it seems reasonable to suggest that previous, earlier work in this field, which has often been of poor methodological quality, may have been somewhat misleading.

3. UK Education Production Function Studies

There is considerable interest in the impact of expenditure on educational performance in schools in the UK. Current expenditure on primary and secondary schooling in England alone exceeds £20 billion (DfEE, 2000). Clearly the application of such resources must be justified and schools, LEAs and central agencies held accountable for their use. Recent initiatives by Central Government, such as the Local Management of Schools (1990) and the

⁵³ The author stresses that small class sizes provide only a *potential* explanation for the favourable “P-5” results. It is quite conceivable that other (unobserved) differences between “P-5” schools, and regular and city-wide schools, are determining those results.

Fair Funding arrangements introduced in 1999⁵⁴, have attempted to increase both accountability and value for money in education. However, as the review of the literature presented below shows, even in the UK there is still a paucity of concrete empirical evidence on the effective use of resources in schools, and thus a lack of evidence-informed guidance on how schools can make best use of their available resources.

The UK empirical evidence examined in Section 3 covers education production function studies. These are concerned with the relationship between resources and educational outputs at LEA and school level. Section 4 considers studies that broadly fall into the category of cost-effectiveness, since they deal with aspects of institutional provision and seek to compare outputs from that provision with its costs.

The criteria for the selection of UK research for inclusion in this review have been drawn more widely than for those considered in wider international research, otherwise there would have been a very small number of UK studies included, and thereby some potentially useful evidence might have been excluded, as certainly would many illustrations of the problems of undertaking research in this area. The UK studies included meet all the following criteria:

1. publication in refereed journals or in reports published by government or non-governmental departments, including, for very recent research, ESRC final award reports and papers submitted to journals;
2. use of at least one of the following school output measures: examination attainment, cognitive test scores, continuation at school, drop out, truancy, attendance, or earnings (only in studies also including school level educational attainment);
3. inclusion of estimates of the impact on output measures of any of the following resource/input measures: expenditure per student, pupil-teacher ratio (PTR), class size, teacher costs per student, non-teaching staff costs per student, non-staff costs per student, measures of teacher quality (such as teacher ability, experience and qualifications);
4. inclusion of student prior attainment or family background and personal variables as controls (*i.e.* a value-added specification);
5. have a clearly identified method of estimation;
6. report the magnitude of the effects of input measures on output variables, including significance or standard errors.

The only exception to these criteria is the inclusion of some studies in the cost-effectiveness Section in which outputs and costs are not directly linked. This exception also includes OFSTED evidence on school efficiency and value for money that uses a qualitative framework that meets only criteria 1 to 4.

The UK literature has relatively few methodologically strong studies. It is also patchy and lacks both depth and breadth of coverage with respect to the different phases of education and datasets used. The research has been restricted by the lack of suitable and accessible data, as has been highlighted by Mayston and Jesson (1999). (See Sections 1 and 5 for a

⁵⁴ Local Management of Schools is a system of school funding introduced in the UK from 1990/91. LEAs are required to delegate a large proportion of funding directly to schools through a process of formula funding. The Fair Funding scheme extended this delegation principle, defining those service aspects a LEA could itself provide and those for which funding must be devolved to schools. Schools' budget shares, the formulae, are defined on an objective, needs led basis, with age-weighted pupil funding accounting for at least 80% of a school's budget.

discussion of these data issues.) The limited number of studies and their limited coverage of important school resourcing variables means that, unlike the international research, it is not possible to organise the review of UK evidence around the key resource variables. Instead, the education production function studies have been divided into two types, depending upon the level of aggregation - LEA or school/student level. The first section reviews studies using LEA-level data for both outputs and inputs.

3.1 LEA-level education production function studies

There are five studies reviewed that utilise both output and input data aggregated at LEA-level. The studies and their main features are summarised in Table 3.1.1(a) below. The two linked studies undertaken by the DES⁵⁵ (1983, 1984) are the earliest education production function studies in the UK literature reviewed. The third study by Lord (1984) uses OLS as well as 2SLS regression techniques whereas the fourth (West *et al*, 1999) reports only partial correlation coefficients. The last study considered in this section - Jesson *et al* (1987) - differs from the others in applying DEA. The extent to which these five studies attempted to overcome the main methodological difficulties is indicated in summary form in Table 3.1.1(b).

3.1.1 Regression studies

The 1983 DES study found that socio-economic variables explained between two thirds and three quarters of the variance in exam performance and that the expenditure variable was insignificant or barely significant and wrongly signed. The 1984 DES study which used more recent census data for the socio-economic status (SES) variables, found these explained 10% more of the variance in exam results than in the earlier study. The only evidence found for a positively signed and significant resource input effect was that teacher expenditure per student explained 1 to 3% of the variance in three of the output measures. These were the proportion of maintained school leavers with 1 or more A-level passes, the proportion with five or more higher O level/CSE passes and those with one or more higher O level/CSE passes. Also the PTR was significantly and positively related to the proportion of school leavers with no qualifications.

Lord (1984) found that SES status and poor housing were the most important of the factors explaining educational outcomes (with between 46 to 57% of variance explained). He found no significant relationship between education spending and outcomes, apart from a perverse negative impact of spending per pupil on the proportion of pupils leaving school at the minimum legal age. However, he did find a negative impact of inexperienced teachers on both measures of examination output. In both cases a three-percentage decrease in teacher inexperience would increase exam results by one percentage point. The proportion of graduate teachers also exerted a positive impact on exam performance. A 4% increase in graduate staff would increase the percentage of school leavers with 5+ O/CSE higher grades by one percentage point, while the impact on the 1 or more O level/CSE grades indicator was three times as high. Lord's findings are consistent with the 1984 DES study, which found expenditure on teachers had a positive impact on exam results.

West *et al* (1999) have undertaken a recent study using aggregate LEA expenditure, exam performance and SES data. When reviewed it had not been published and was part of an ESRC End of Award Report. The authors show that LEA expenditure on education is very highly correlated with Standard Spending Assessments (SSA), which in turn depend on pupil numbers and the Additional Educational Needs indicators. Thus, in this dataset education expenditure is linked to SES indicators and this gives rise to the usual endogeneity

⁵⁵ The Department of Education and Science (DES) became the Department for Education and Employment (DfEE) in 1995.

problem. The study does not satisfactorily indicate how the endogeneity problem is tackled through the statistical methods employed. The authors find that KS1 test results and GCSE examination results at LEA-level are inversely and strongly correlated with indicators of social deprivation. Positive and significant partial correlations of 0.3 and 0.41 between expenditure per pupil and the two GCSE exam result indicators after controlling for SES variables are also reported. Unfortunately, in the absence of reported multiple regression estimates and indications of how endogeneity is corrected for, it is difficult to assess these statistical results. Consequently, no precise conclusions can be drawn from this work of the impact of expenditure on pupil outcomes.

Table 3.1.1(a) LEA-level education production function studies (regression): main features

Authors	Output measure	School quality variables	Controls	Data	Statistical technique
DES (1983)	Proportion of maintained school leavers achieving: 1 or more A-level passes; 5 or more higher grade passes at O level (A-C) or CSE (grade 1); 1 or more higher grade passes at O level (A-C) or CSE (grade 1); no graded passes at O level (A-C) or CSE (grade 1).	Secondary school expenditure per pupil	Children: 1) born outside UK or non-white; 2) living in household with head unskilled/semi-skilled manual worker; 3) in households lacking amenities; 4.) living in density > 1.5 per room; 5.) in 1 parent families; 6.) in families with 4 or more children. LEA-level: free school meals; SES measure; unemployment; population.	School Leavers Survey 1977/78 to 1980/81 (mixed exam year cohorts). 1971 population census. LEA education statistics returns.	Stepwise OLS
DES (1984)	As above plus: 6 or more graded results at O level/CSE; 2. or less graded results at O level/CSE.	SECONDARY TEACHING EXPENDITURE PER PUPIL; Secondary non-teaching expenditure per pupil; PTR for 11-16 age group in 1983; PTR for 16-18 year olds in 1983.	As above.	As above except 1981 census data.	Stepwise OLS
Lord (1984)	5 or more higher grade passes at O level (A-C) or CSE (grade 1) 1 or more higher grade passes at O level (A-C) or CSE (grade 1) Proportion not continuing in post-compulsory education Delinquency rate of under 18s.	Secondary teachers with degree Teacher experience Class contact ratio Pupil-teacher ratio Expenditure per pupil Secondary expenditure per pupil	DES variables 1 to 6 above. Percentage of pupils in LEA attending independent schools	As above plus additional data from DES (Form 7, teachers' pension). Criminal Statistics.	OLS 2SLS
West et al. (1999)	% of pupils attaining: KS1 level 2 plus in 1996; 1+ A*-G GCSE passes 1996, 1997; 5+ A*-C GCSE passes 1996, 1997.	Total education expenditure per pupil	Percentage of children: from outside UK, Ireland, USA and Old Commonwealth; in lone parent families; dependent on income support; with statemented/non-statemented SEN.	SSAs & LEA expenditure outturn 1994/5 to 1997/98. Form 7 1996. DfEE NC test/ exam tables.	Partial correlation, after controlling for income support and AEN separately.

Table 3.1.1(b): LEA-level education production function studies: methodology

Authors	Methodological issue		
	1.1.1 Endogeneity	1.1.2b Functional form	Omitted variable bias
DES (1983)	✗	✗	✗ /4
DES (1984)	✗	✗	✗ /4
Lord (1984)	4	✗	✗ /4
West et al. (1999)	✗	None	✗ /4
Jesson et al. (1987)	✗ /4	✗/4	Caveats on sensitivity of DEA to inclusion of different variables

Key: 4 Attempt to overcome methodological difficulty; ✗ No attempt to overcome methodological difficulty; ✗ /4 Some attempt to overcome methodological difficulty.

3.1.2 Data Envelopment Analysis Study: Jesson, Mayston and Smith (1987)

Using the same data as the 1984 DES study, Jesson *et al* (1987) apply data envelopment analysis (DEA) (Section 1.2.2). DEA is a useful antidote to single output education production functions since it properly treats efficiency as a relative concept. A school's efficiency is measured relative to that of its peer group with the same linear combination of outputs. Two output measures were used by Jesson *et al*:

Output 1: percentage of children getting 5 or more GCE O levels or CSE grade 1 passes;

Output 2 percentage of children getting 3 or more graded passes at CSE or GCE O level.

The input variables were:

percentage of children in LEA area whose head of household is a non-manual worker;

percentage of children not from one parent families;

percentage of children born in UK, Ireland, USA or Old Commonwealth or whose heads of household were born in these countries;

secondary school expenditure per pupil, including both teaching and non-teaching costs.

LEAs can be judged efficient for different combinations of the two outputs, provided any particular combination of the outputs is not produced using more of at least one 'input' than any other LEA producing that combination of outputs. This study indicated that 27 out of 96 LEAs were 'inefficient'. The authors warn, as indeed we have highlighted in section 1, that DEA is limited by the adequacy of the data and is sensitive to the outputs and inputs chosen for inclusion. Hence they regard the study as illustrative not definitive. However, the use of DEA in this study is important, since it serves to remind us that if there are two or more LEA education outputs which are substitutes for each other, then single output-expenditure measures cannot be used to judge LEA efficiency.

3.2 School level education production function studies

The school level education production function studies considered fall into three categories. First, there are studies using student level data, which all derive from the National Child Development Survey dataset. Second, are studies using data aggregated at school level from the Annual School Census (Form 7)⁵⁶, school examination performance and general population census datasets. Finally a review of UK class size studies is referred to: there are few of these large-scale quantitative studies and all were undertaken 12 or more years ago.

3.2.1 National Child Development Survey studies

The most recent and most econometrically sophisticated studies are to be found in four papers utilising National Child Development Survey (NCDS)⁵⁷ data. The studies reviewed are: Dearden *et al* (1997), Dustmann *et al* (1998), Dolton and Vignoles (1999), and Feinstein and Symons (1999).

All three studies estimate an education production function equation of the form:

$$O_i = f(T_i, F_i, S_i, P_i, E_i) \quad (3.1)$$

where:

T_i = prior attainment of individual i ;

F_i = Family background variables,

S_i = Schooling inputs or school quality variables;

P_i = Peer group effects;

E_i = local environmental effects

The main features of these four studies are summarised in Table 3.2.1(a) and an assessment of their methodology is given in Table 3.2.1(b). The NCDS provides individual level data on educational outcomes, prior attainment at 7 and 11, family background, and school quality, in particular type of school attended, its pupil-teacher ratio and the child's class size at 16 for maths and English. The NCDS data are supplemented by other data on expenditure and resourcing at LEA-level (from LEA returns) and by census data on SES variables at LEA-level.

Outcome variables

All four studies use examination results as a measure of school output, although the exam measures derived differ, as shown in Table 3.2.1(a). Dolton's and Vignoles' maths and English exam scores, Feinstein's and Symons' English attainment measures, and Dustmann *et al*'s number of exam grades, correspond directly to the qualifications achieved by pupils at the age of 16. The other measures of examination attainment combine at least two stages of schooling, though a single year's pupil-teacher ratio (PTR) is used as a regressor.

Dearden *et al* and Dolton and Vignoles also estimate earnings equations as a function of qualifications (*i.e.* highest previous educational attainment), family and individual

⁵⁶ The Annual School Census (Form 7) collects information from each school every year. Data collected include pupil and teacher numbers, as well as certain socio-economic information. Data are collected at school level. However a pupil level Annual School Census is currently being piloted. This will be fully operational in 2002 and will be combined with pupil-level SAT/GCSE scores to form the National Pupil Database.

⁵⁷ The NCDS is a longitudinal study which includes all children born between 3 and 9 March 1958. They were surveyed at birth and then at ages 7, 11, 16, 23 and 33.

characteristics and school quality variables (including LEA-level inputs). Dustmann *et al* include career choice as a school outcome variable, where the choice is between staying on at school, training (full- or part-time) or entry into the labour market. Only Feinstein and Symons include peer group effects.

Modelling

The studies are also notable for their concern with correcting and checking for bias due to omitted variables and endogeneity (Section 1.1.1) as shown in Table 3.2.1(b). They all attempt to overcome the endogeneity problem via the ‘saturation’ method, which involves testing a sequence of model specifications. The initial model contains a limited number of variables. In subsequent models more and more variables are included to show the effects of controlling for different measures of ability/prior attainment and family background, local environment and school quality variables. Feinstein and Symons also supplement OLS estimates with 2SLS and sensitivity analysis which indicates that the endogeneity bias is unlikely to be serious. Dustmann *et al* include a simultaneous equation testing of their career choice model, while Dolton and Vignoles report a test with an instrumental variable.

Dearden *et al* try an additional functional form. They extend the basic model – equation 1 above – by introducing interaction terms: the pupil-teacher ratio is interacted with school type and ability.

Findings

The findings, summarised in Tables 3.2.1(c) and 3.2.1(d) focus in on the school quality variables and report coefficients on these variables estimated after including the largest range of control variables. The studies confirm the overwhelming importance of prior attainment/ability and family background variables in determining school educational attainment.

No significant relationship between school input variables and labour market outcomes was reported by Dolton and Vignoles, even after trying a variety of specifications (*e.g.* including only comprehensive school pupils). They also sought to overcome endogeneity by using as an instrumental variable (Section 1.1.1c) the random variation in educational resourcing levels that followed a change in LEA organisation in 1974. However, Dearden *et al* found some school quality variables to be significant and correctly signed in wage equations (see Table 3.2.1(d)). This finding came out of specifying interactions between the pupil-teacher ratio (PTR) and school type and PTR and ability. The PTR did have significant and negative effects for men who attended secondary modern schools, and lower ability women. Thus the Dearden *et al* study indicates the importance of model specification and the use of interaction terms (Section 1.1.1b) to probe how resources may be differentially effective for different types of student.

Table 3.2.1(a) Student level education production function studies: main features

Authors	Output measure	School quality variables	Controls	Data	Statistical technique
Feinstein and Symons (1999)	EXAMS 1. English: highest grade attained in national exams in English up to age of 21 (has 8 categories). 2. Mathematical ability at 16 measured by NCDS test. 3. Index of overall exam performance in all subjects.	School type (single sex, private, grammar, comprehensive, technical, secondary modern) PTR at school level. Pupil in top stream Not in top stream of streamed classes	<u>PRIOR ATTAINMENT:</u> NCDS 'ability' tests in maths at 11 <u>FAMILY BACKGROUND:</u> Parent interest variable; Father in top or middle SES; Father and mother stayed on at school; No of older and younger siblings; Father plays role in upbringing. <u>PEER GROUP</u> composite variable made up of: % of children in class with fathers in non-manual occupations; % of children in class only taking GCE exams; % of children in class only taking CSE exams; % of children in previous year's class who stayed on in education. <u>ENVIRONMENT</u> local unemployment rate, % of unskilled manual workers	NCDS General population census	OLS and 2SLS Monte Carlo simulations: sensitivity analysis to check effect of low correlation in 2SLS between instruments and endogenous variables.
Dustmann et al. (1998)	EXAMS Number of O level and CSE grade 1 passes DESTINATIONS Career choice (staying on, full & part-time training, labour market)	School type PTR at school level	<u>PRIOR ATTAINMENT</u> NCDS 'ability' tests in maths and English at 7 & 11. <u>FAMILY BACKGROUND</u> Family income; parents working; Parents' education; Child has separate room; No of older and younger siblings; Parental interest; Parents want child to sit A-levels / go to university. <u>ENVIRONMENT</u> Local unemployment rate; % of unskilled manual workers	NCDS General population census	OLS (ordered probit) Varied specifications. 2 stage estimates with instrumental variables.
Dolton and Vignoles (1999)	EXAMS English and maths exam results at age 16:	School type English and maths class	<u>PRIOR ATTAINMENT</u> NCDS 'ability' tests in maths and English at 11. <u>FAMILY BACKGROUND</u>	NCDS LEA education	OLS (ordered logit)

	<p>1 = no qualifications in E or M; 2 = unclassified grade; 3 = CSE grade 3 or 4; 4 = O level grade D or E; 5 = O level C or CSE 1; 6 = O level grade A or B</p> <p>WAGES Log of gross hourly pay at age 33 for employed males.</p>	<p>sizes (at 16); Square of class size; PTR at school level; School size; % students staying on; LEA expenditure per pupil; Teachers' salaries per pupil; PTR at LEA -level; Child setted or streamed.</p>	<p>Gender; race; social class; Home ownership; Number of. siblings; Father present; Parental attitude to staying on at school.</p> <p><u>PEER GROUP</u> Attended school where: <20% pupils non-manual; >80% pupils non-manual.</p>	<p>statistics</p>	<p>Varied specifications.</p>
<p>Dearden et al. (1997)</p>	<p>EXAMS Highest qualification obtained at school (A-level, 5+O-level A-C or CSE 1, 1+ O level A-C or CSE 1, CSE 2-5, none). Highest educational qualification obtained at age 23 or 33. WAGES Wages at 23 and 33: hourly real gross wage rate in 1995 prices (of those in employment in 1981 and 1991)</p>	<p>School type PTR at child's school at 11 & 16 LEA expenditure per pupil LEA average teacher salaries in primary and secondary schools in 1969 and 1974</p>	<p><u>PRIOR ATTAINMENT</u> NCDS 'ability' tests at 7 in verbal and maths ability. <u>FAMILY BACKGROUND</u> Parental interest; Father's social class; Father's and mother's education; In receipt of FSM; Family financial difficulties; No. of siblings and No. of older siblings. <u>ENVIRONMENT</u> Regional school dummies (11); Census (1971) SES variables of enumeration district in which child lived; Social deprivation level of LEA (1971); Size of LEA and its spending needs.</p>	<p>NCDS LEA education statistics General population census</p>	<p>OLS (ordered probit) Varied specifications, including interaction terms between PTR and school type; and PTR and ability.</p>

Table 3.2.1(b): Student level education production function studies: methodology

Authors	Methodological issue			
	1.1.1 Endogeneity	1.1.2a Aggregation Bias	1.1.2b Functional form	Omitted variable bias
Feinstein and Symons (1999)	4	4	×	4
Dustmann et al. (1998)	4	4	×	4
Dolton and Vignoles (1999)	4	4	×	4
Dearden et al. (1997)	4	4	4	4

Key: 4 Attempt to overcome methodological difficulty; × No attempt to overcome methodological difficulty; × /4 Some attempt to overcome methodological difficulty.

Table 3.2.1(c): Summary of findings of student level studies on effect of resource variables on exam results

	Feinstein and Symons (1999)	Dustmann et al. (1998)	Dolton and Vignoles (1999)	Dearden et al. (1997)
Class size	Not included	Not included	Significant but positively signed. Square of class size significant negatively signed.	Not included.
School PTR	Insignificant	Significant when school type not included. Insignificant once school type included.	Significant: Maths score coefficient = -0.091; English score coefficient = -0.068	Not significant except for negative effect on men attending secondary moderns and lower ability women.
LEA PTR	Not included	Not included	Insignificant	Not included
LEA exp. per pupil	Not included	Not included	Insignificant	Insignificant
School type	Compared to comprehensive, coefficients for all exams is: grammar = 7.56 sec. mod. = -2.32 private is insignificant. Peer group = 10.29 Top stream = 7.57 Not top stream = -5.42	Significant. Compared to secondary modern coefficient on exam score is: Private = 2.087 Grammar = 1.916 Technical = 1.137 Comprehensive = 0.69	Maths/English : coefficients compared to comprehensive: private = 1.168(M): 0.882(E) grammar = 0.585(M): 0.886 (E) secondary modern = -0.076(M): -0.193(E)	Men: grammar and private school attendance at 16 has significant and positive effect, secondary modern negative effect. Women: girls school had significant positive effect.

Table 3.2.1 (d): Summary of findings of student level studies on effect of resource variables on other outputs⁵⁸

OUTPUT	Dustman et al. (1998) CAREER CHOICE	Dolton & Vignoles (1999) WAGES AT 33	Dearden et al. (1997) WAGES AT 23 & 33
Class size	Not included	Insignificant in full specification	Not included
School PTR	Significant. decrease in PTR by 1 st. dev. (2.3) increases probability of staying on at school by 6-7 percentage points.	Not significant once control for ability, qualifications, personal factors and experience.	Negative and significant effect on wages at 33 for women attending grammar schools.
LEA PTR	Not included	Insignificant	Not included
LEA spending per pupil	Not included	Insignificant	10% inc in average secondary teachers' salary per pupil leads to 10% higher male wages at 23. 10% increase in LEA secondary school expenditure per pupil leads to 3.1% higher female wages at 23.
School type	Grammar/private school increases staying on by 16/19 percentage points.	Significant.	Private school has 9% impact on male wages at 33. Has positive impact on female wages at 33.

Dustmann *et al* also report significant effects on career choice of the school level PTR after controlling for school type. They find that a lower PTR increases the probability of a student deciding to stay on at school after 16. The PTR also has a significant effect on exam results, though this becomes insignificant when school type is included. This is because PTR and school type are quite highly correlated, so PTR acts as a proxy for school type. The authors note that in the absence of school types in the exam equation, PTR explains less than half the impact on exams than does school type, indicating omitted variables, such as peer group effects and teacher quality.

Dolton and Vignoles report that the PTR is significant and negatively signed in their exam equations, which include school type, while Feinstein and Symons find it insignificant. However, the measures of attainment used by Dolton and Vignoles are a more valid construct for school output than those of Feinstein and Symons; the former use English and Maths exam results and all subject exam results taken at 16, whereas the latter use the NCDS test of maths and measures of highest attained qualification from school.

Only Dolton and Vignoles include class size as a regressor in school attainment equations, where it was significant but positive, probably because of its association with pupil ability. (It became insignificant when class size squared was added to the model). Feinstein's and Symons' study is useful in indicating the importance of the peer group in explaining attainment at school. None of the studies explored the relationship between peer group and class size.

In summary, three of the studies provide evidence of the effects of school resource variables on school attainment and one of them of effects on wages. By progressively adding in more variables these studies go some way in controlling for endogeneity. Fewer school quality variables are reported as significant when a larger number of explanatory variables are controlled for, indicating the likelihood of omitted variables bias in other studies which use only a few control variables. The NCDS studies have also tackled the endogeneity problem by means of instrumental variables. Inclusion of interaction terms enables the differential effects of resources for different types of student to be explored. Consequently,

⁵⁸ Note that Feinstein and Symons only use exam scores as their output. Consequently they are not included in this table.

some positive effects of school quality variables on educational outcomes are now being discovered using UK data.

3.2.2 Studies using data aggregated at school level

Another small group of studies utilises input and output data at school level. The one most pertinent to this review is Bradley and Taylor (1998). The main features of the study are summarised in Table 3.2.2.

Table 3.2.2: School level study (Bradley and Taylor): main features

Output measures	Proportion of a school's students aged 15 in the year prior to taking GCSEs who obtained 5 or more GCSE grades A* to C
School quality variables	% of students taking Advanced level courses; % of students taking vocational courses; % of teachers with formal teaching qualifications; part time/full time teacher ratio; school type (LEA or GM; single sex); number of pupils in the school.
Controls	% of school's students: % entitled to free school meals; % with special needs; % from non-white backgrounds.
Data	DfEE Schools Performance Tables: all non-selective English state secondary schools 1992 to 1996. Annual School Census (Form 7)
Statistical technique	OLS estimates of ordered logit equations of exam results estimated separately for 11-16 and 11-18 schools in 1992 and 1996. Endogeneity of the FSM variable is checked for by use of instrumental variables.

Not surprisingly, the strongest relationship found was between exam results and free school meals. Other persistently strong and positive effects for all schools were:

- being a voluntary aided school;
- being a girls only school
- the proportion of part-time teachers
- school size (number of students) (positive coefficient on number of pupils, negative on number of pupils squared).

There were positive significant effects for 11-18 schools from:

- the proportion of students taking A-level/ vocational courses
- being GM in 1996.

Being a secondary modern school had a negative impact on exam results. The only internal school resource variable which was significantly and positively related to exam performance was the proportion of part-time teachers, possibly because this indicates greater staffing flexibility and better matching of teacher strengths to curriculum requirements. The pupil-teacher ratio was insignificant, except in 1992 for 11-18 schools, where the sign on the coefficient was positive (*i.e.* opposite to that expected). The authors

explain this finding as due to lower PTRs being associated with higher proportions of low ability students (endogeneity). School size exerted a positive effect on exam performance. From the estimated coefficients the optimal size of school was modelled. Exam performance was maximised at the size of 1200 for 11-16 and at 1500 for 11-18 schools. In comparison, in 1996 the actual mean size for 11-16 schools was 765 and for 11-18 schools was 1010. Roughly 70% of schools are below optimum size with respect to this criterion.

A second study involving Bradley *et al* (1999), utilising the same database with additional census data and measures of proximity of other schools, is a DEA study in which free school meals and the proportion of unqualified teachers are treated as the only two inputs in the education production function for estimating the efficiency frontier and departures from it. The study does not add any new information concerning the effects of school inputs in an education production function to that given in Bradley and Taylor (1998).

3.2.3 UK class size studies review

The most recent review of UK class size research found is Blatchford and Mortimore, (1994). Blatchford's and Mortimore's review of UK class size studies is quite brief as it is part of a wider ranging article. Four of the studies they review (Morris, 1959; Wiseman, 1967; Davie *et al*, 1972; and Little *et al*, 1973) are correlational and large scale, being described as well designed and well carried out. These studies tested for associations between class size and pupil attainment. In general they found that pupils in larger classes did better than those in smaller classes. Attempts were made to control for endogeneity by including variables such as parental occupation (Wiseman), parental interest and occupation (Davie's study using NCDS data) but the advantage of larger classes remained. ORACLE (Galton and Simon, 1980), a 'large scale' study using classroom observation data, found that while class size had some association with classroom interactions, larger classes did not result in lower rates of pupil progress. The only UK study with any contrary findings is the Junior Years Study (Mortimore *et al*, 1988) which reported that for a sample of 50 London primary schools smaller classes were associated with greater progress for 8 year olds in mathematics and non-cognitive development.

Blatchford and Mortimore note that no UK studies have used experimental methods. They conclude, having also reviewed international research, that there is now firm evidence of a link between lower class size and higher educational attainment but only in the early years, particularly for socially disadvantaged children and only for classes smaller than 20. This conclusion, however, rests almost entirely on the international evidence. The reviewers also comment on the lack of recent research on class size in the UK. Currently Blatchford and Goldstein are undertaking a large-scale study using MLM but they have not yet produced published findings.

3.2.4 Conclusion

Education production function research in the UK has been severely hampered by the lack of good quality data. Apart from the NCDS studies, all the others suffer from the aggregation problem since they use LEA or school level data. Furthermore, LEA studies have not proved a particularly fruitful line of inquiry for the education production function because of the endogeneity problem (Section 1.1.1). The data are also simply at too high A-level of aggregation for testing the important relationships between school level variables and outcomes. Where any positive impacts of resources on outputs have been found in LEA and school aggregate level studies, these concern teacher quality variables and not overall expenditure per pupil or PTR.

More recent studies utilising student level data from the NCDS, together with school and LEA-level resource input data, have been more successful in controlling for endogeneity and detecting some school resource effects. The larger range of variables has enabled these studies to make progress in reducing omitted variables and endogeneity bias. They have also used more sophisticated and varied model specifications. These studies have produced some evidence of school quality variables impacting positively on non-exam outcomes. However, these studies have not always utilised measures of school output with high construct validity and have had recourse to only limited data for school level resource utilisation and for instruments for controlling for the endogeneity of the school resource variables. The interaction effects in Dearden *et al* are indicative of the potential importance of differential resource effects for students according to gender and ability, which studies utilising school level data would miss.

4. UK Cost Effectiveness Studies of School Provision

As well as the application of statistical techniques to analyse the relationship between expenditure and school performance, other methodologies have also been applied to explore links between resource inputs and performance outputs⁵⁹. This section describes the application of some of these techniques in a UK context. The studies included fall into three groups:

- i) comparative analysis of course costs and performance on an inter-organisational basis. This work is largely restricted to costing provision for 16-19 year olds.
- ii) The evaluation of specific programmed education interventions, comparing their costs and their effects on pupil performance.
- iii) The application of the OFSTED inspection framework to provide comparative, judgemental evaluations of schools' use of resources to deliver pupil performance.

4.1 The cost-effectiveness of A/AS level provision

Analysis of A-level provision has been a promising line for researchers to pursue because of the availability of student level prior attainment data, namely GCSE results. The research on A-level cost effectiveness falls into two types. The first is a set of related academic studies of the cost-effectiveness of A-level courses in 12 institutions representing a range of providers. The second set consists of studies sponsored by governmental bodies (Audit Commission, OFSTED, DfEE), either on value-added A-level analysis or on the cost-analysis of A-level provision in different institutions.

4.1.1 Fielding and Thomas: cost-effectiveness of A-level provision in 12 institutions

Two important studies of the cost-effectiveness of A-level provision were undertaken by Fielding (1995 and 1998), utilising data collected by Thomas from primary sources (Thomas, 1990). The Fielding studies reanalysed Thomas' data using multi-level modelling. Fielding's 1998 study was concerned with institutional costs, whereas the 1995 study also reported cost-effectiveness estimates by institution type for social costs (institutional plus students' private costs of approximate earnings foregone). The two studies reached broadly

⁵⁹ In addition to the material reviewed here, there is huge body of literature relating to school effectiveness, which has also looked at cost effectiveness issues. A recent summary of this literature can be found in Teddle and Reynolds (2000), and hence, whilst we sometimes allude to this literature, we do not explicitly include this research in our review.

the same conclusions about relative cost-effectiveness, so this review focuses on the 1998 study.

Data were collected for three cohorts of examinees in 1980, 1981 and 1982 for 1162 teaching groups with 10685 students in 12 institutions (6 school sixth forms (SF), 3 sixth form colleges (SFC), 1 tertiary college (T) and 2 FE colleges). Output was the A-level score by subject at student level, with prior attainment measured by O level/CSE scores.

Costs were allocated to each teaching group by A-level subject. These consisted of:

- institutional overhead costs for each year 1978 to 1981; expressed in 1981 prices;
- cost of time-tabled teacher time per teaching group.

The dependent variable was the cost-effectiveness ratio obtained by dividing the student subject A-level score by the cost per student. Because there were only 12 institutions, the main model reported was fitted at two levels (student and teaching group) with each of the 12 institutions represented by a dummy variable. The individual students' O-level score was used as a fixed control variable at level 1 and the mean teaching group O-level score as a fixed variable at level 2. Additional variables included were number of teachers who taught the group, number of candidates in the group and dummy variables for 7 types of A-level subject.

83% of the explained variation in the cost-effectiveness ratio (Fielding, 1998) was accounted for at the student level and 17% at the teaching group level. Six of the seven subject dummies were significant. O-level score was highly significant. The number of candidates (reflecting class size) was positive and significant, indicating, that larger classes at A-level are more cost-effective.

All except two of the institutional dummy variables were significant. The institutional dummies were interpreted as indicating a ranking in institutional types, with sixth form colleges the most cost-effective and school sixth forms the least cost effective, corroborating Thomas' (1990) findings. An odd result, which is not commented upon, is that cost-effectiveness was significantly and negatively related to the teaching group's mean O-level score. This may well be due to endogeneity, in that school sixth forms have higher average prior attainment but smaller teaching groups than FE colleges.

The small number of institutions included in these studies, clearly limits their general applicability. Larger and more comprehensive data sets have since become available and will enable more extensive research to be carried out on these issues. Future research based on these larger data sets should be more generalisable.

4.1.2 Officially sponsored reports on A-level cost-effectiveness

The DfEE, Audit Commission and OFSTED have commissioned and undertaken a number of research studies into the cost-effectiveness of A-level provision in different institutions. Some draw on national data sets and others utilise fieldwork data from a relatively small sample of institutions. The studies included are summarised in Table 4.1.2(a).

Table 4.1.2(a): Officially sponsored reports on A-level cost-effectiveness

Study	Type of study	Data source	Main input variables	Main relevant findings
<i>Unfinished Business</i> (Audit Commission and OFSTED, 1993)	Costs and value added A-level results	Field work: 42 schools and colleges in England and Wales, 1991.	GCSE Unit cost of A-level courses (teaching and non-teaching costs).	No relationship between A-level value added and unit costs.
<i>Effective Sixth Forms</i> (OFSTED, 1996)	Costs and HMI judgement of quality of provision	Field work in 92 school sixth forms in 30 LEAs, 1993/94. Case studies of 18 6 th forms, 1994/95.	Cost of sixth form in comparison with budget received for 6 th form.	About two thirds of sixth forms were cost effective. Difficult for small sixth forms to be cost-effective without consortia.
<i>Two B's or Not</i> (Audit Commission, 1993)	Value added: A-level points score. OLS and MLM	Student level: 1721 candidates, 1988, from YCS.	GCSE, gender, institution type, parents' education, social class, ethnicity.	37-39% A-level variance explained. Significant: GCSE, female and non graduate parents (both negative). <i>No significant difference for type of institution attended.</i>
<i>Value Added for 16-18 Year Olds in England</i> (O'Donoghue et al., 1997)	Value added: A-level points score. MLM.	Student level. Bath examinations data for 3 cohorts, 1993-95. 504680 candidates in 2824 institutions.	GCSE, gender, institution, institution type	55% A-level variance explained. Significant: GCSE, female (-ve), selective school (+ve), GCSE score of year group (+ve), school FSM% (-ve). <i>Institutional type: comprehensives & FE colleges –ve effect compared to 6th form colleges; latter similar to GM & grammar schs.</i>
Public Funding Costs of Education and Training of 16-19 Year Olds (DfEE, 1998b). Developing Funding Cost Comparisons (DfEE, 1998a)	Public funding costs of different types of institution	FEFC funding tariffs, CIPFA schools finance data, Form 7, YCS (for course retention), Bath exam data.	Unit cost per successful completion of 2 and 3 A-level packages by institution type.	Unit cost ranking from highest cost: GM schools; LEA schools; 6 th form colleges; FE general colleges.

Unfinished Business (Audit Commission and OFSTED, 1993) is the only report of those reviewed which directly compared the cost of A-level provision at institutional level with value-added A-level scores. To estimate the unit costs of A-level provision at institutional level, four A-level courses at each of the 42 institutions studied were costed. The methodology for costing teacher staffing required data on the number of course hours a week taught by teachers. This was put on an annual basis and multiplied by the cost of an hour's teaching⁶⁰ and divided by student numbers on the course. Non-teaching recurrent costs⁶¹ were apportioned according to course teaching hours. Unit course costs per A-level completion at each institution were compared with value-added A-level subject scores: no significant correlation between the two variables was found.

The Audit Commission's (1993) study found that while schools tended to have higher ability candidates than further education colleges, the data did not show any single type of institution as consistently more effective at A-level than the others, when full account was taken of differences in examination qualifications of their intakes at 16 plus. They all 'added value' to GCSE in roughly equal measure.

OFSTED's (1996) study of *Effective Sixth Forms* did not cost courses directly. Instead, it compared the percentage of staffing allocated to sixth form teaching in a school with the percentage of the budget generated by sixth form Age Weighted Pupil Units. If this did not exceed 100% and the school was judged by the HMIs to provide an effective education for its students, then the sixth form was deemed cost-effective. Using this criterion, two thirds of sixth forms were judged to be cost-effective.

Both the studies reviewed above, which attempted to link the effectiveness of A-level courses at institutional level with unit costs of the courses, gathered data by fieldwork in a relatively small sample of institutions. Some later work, reviewed below, used larger datasets and reported findings on the comparative educational effectiveness of institutions or on comparative funding.

Three interrelated studies (DfEE 1998a, 1998b and O'Donoghue *et al* 1997) were jointly commissioned by the DfEE. The first two studies were concerned with costing the funding of post-16 qualifications provided by the different types of institution, while O'Donoghue *et al* was a supporting multi-level analysis of A-level results. Though not able to bring A-level output and cost measures into a single regression, the three studies together provide evidence on the cost-effectiveness of different types of provider.

The DfEE (1998a and 1998b) studies calculated the public funding costs per qualification package of the different sectors providing equivalent post-16 qualifications. The studies made clear that the public funding costs of completed 2 and 3 A-level packages were estimates not actual expenditure costs as 'it is not possible to provide expenditure-based figures' (DfEE, 1998b, p.2). The costing methodology adopted was driven by the need to make school funding for A-level packages equivalent to that received by FE colleges. These are funded by a unit tariff calibrated by course and translated into cash according to an exchange rate called the Average Level of Funding (ALF) that still differs by FE college. FE costs of provision were estimated according to the ALF of the median college. The FE tariff unit methodology was then replicated for schools using cost data from Section 122 returns, Form 7 (for funded pupil numbers) and Bath exam data (for completed qualifications by institution). Youth Cohort Survey data were used to provide an estimate of drop out and retention rates for each sector. Unit costs were calculated for qualifications completed, while

⁶⁰ The cost per teacher hour used in calculating the direct costs of courses depends on assumptions about many hours a week/year teachers normally teach and on how to apportion teachers' non-contact time between work done to support their course teaching and other activities carried out in non-contact time. ⁶¹ Support staff, consumables and equipment, maintenance, administration and other costs.

funding was related to students enrolled. The resulting calculations are shown in Table 4.1.2(b).

Table 4.1.2(b): Public funding costs of 2 and 3 A-level packages: 1996-97

SECTOR	3 A-LEVELS		2 A-LEVELS	
	Cost £	Cost index	Cost £	Cost index
LEA maintained schools	7380	100	4410	100
GM schools	7630	103	4560	104
General FE colleges	6250	85	2980	68
Sixth form colleges	5910	80	3270	74

Source: DfEE (1998b), pp.6 and 10.

Evidence on the value-added outputs of different types of institution for A-levels was provided by O'Donoghue *et al* (1997). This study utilised a database of over 50,000 A-level candidates (see Table 4.1.2(a)) and various model specifications. It found that students in grammar and GM schools and sixth form colleges made similar progress. Students in LEA comprehensive schools made on average three-quarters of an A-level point less progress and students in general FE colleges made 2.5 less A-level points progress than those in sixth form colleges. However, the measure of FE output was biased downwards because it did not take account of vocational qualifications.

A further finding of O'Donoghue *et al* was that there was a small compositional effect. In particular A-level progress was positively related to the average prior attainment of an institution's students and negatively related to the percentage of students entitled to free school meals. Hence slower progress in comprehensive schools and FE colleges was in some part due to peer group effects. The study also estimated that larger group size was associated with a small increase in progress and that changing institution between GCSE and A-level had a very small, almost negligible, negative effect on performance. On the basis of the costing data and findings of O'Donoghue *et al* it was concluded that:

"It is not possible on current evidence to draw firm conclusions about the links between value added and funding" (DfEE, 1998a, p.15).

4.1.3 Summary

To summarise, the overall conclusions regarding the relative cost-effectiveness for A-levels of the different types of institution are still somewhat inconclusive. While general FE sector provision is cheaper than the schools sector, its relative cost-effectiveness in terms of value added compared to the other types of provision has not been fully demonstrated. Differences in unit cost at A-level are largely explained by differences in teaching group size and there is no evidence that increasing group size within the usual size range of 20 or below reduces student progress. Thus the evidence suggests that improving the overall cost-effectiveness of A-level provision depends on structural decisions about rationalising the number of providers, in particular closing small sixth forms and hence increasing the average size of A-level teaching groups.

The methodologies used to investigate A-level cost-effectiveness have been limited by the available data and by the relatively small sample sizes used in most studies. Many studies have therefore not been able to fully address key methodological issues, such as the endogeneity of resource levels. Hence the contribution of this particular literature to the

more mainstream school resourcing research is quite limited. Certainly the A-level cost-effectiveness studies do not provide support for the view that higher resource levels will automatically generate better outcomes.

4.2 Educational interventions

A number of policies and interventions have been devised with the aim of improving performance amongst target sets of pupils. The evidence on two types of intervention – early literacy programmes and special needs provision – is described below. These include two different early literacy projects and one review of a project focusing on children with mild learning difficulties. Many of the interventions have been of an experimental or quasi-experimental nature. Cost-effectiveness approaches have been used to evaluate a number of these interventions.

4.2.1 Reading recovery and phonological intervention

Reading Recovery and alternative Phonological Intervention was a project aimed at improving the reading skills of children who had made a slow start in reading (Qualifications and Curriculum Authority, 1998). It involved 390 six year-old children from 63 schools in London. In September 1992 the children were placed in three groups, matched for initial reading ability for the two terms spanning 1992 and 1993:

- 95 received the Reading Recovery (RR) programme;
- 97 received Phonological Intervention (PI);
- a control group (CG) of 198 children was established, half in the same schools as the RR and PI children and the rest at 18 control schools with no RR or PI programmes.

Reading Recovery was the most resource intensive option as it involved a daily 30 minutes individual session with a RR teacher for about 20 weeks (*i.e.* about 50 hours teacher time per child). Phonological intervention comprised 40 ten-minute sessions over 2 terms (6 hours 40 minutes teacher time per child)⁶². The control group children had their normal programme⁶³.

The children were pre-tested in September 1992⁶⁴. Post-tests were administered in the summer of 1993 (post-test 1) and in the summer of 1994 (post-test 2). There was a further follow up test in autumn 1996 when 342 children were still in the study. In addition, reading test and other data were collected in 1996 from 1398 classmates of the original cohort children.⁶⁵

The effects of RR and PI were estimated using multiple regression to control for the slightly higher initial reading ability of the control group children. Initially the Reading Recovery (RR) children made significantly greater progress than the control group. After one year the RR children were not significantly ahead of the control group in their own school, but were ahead of children in control group schools (those without any RR or PI intervention). Four years later, the RR children had held on to their absolute gains but the difference between them and the control group was no longer statistically significant. In contrast, the immediate post-intervention gains from PI were confined to phonological awareness and had no impact on reading. But by the end of second year, the PI children's

⁶² It should be noted that PI was delivered by the research team, so this is a self-assessed intervention.

⁶³ This often included some support.

⁶⁴ The tests were the British Ability Scale Word Reading Test and Neale Analysis of Reading Ability, Clay's Diagnostic Survey, and the Oddities test for phonological awareness.

⁶⁵ These were tests of children's post intervention reading using BAS word reading and spelling tests, the Neale prose reading test and Oddities test for phonological awareness.

reading and spelling improved significantly. After four years the effect of PI on reading just missed statistical significance, but there was a lasting impact on spelling for all children included in the project, except for children unable to read at 6.

Table 4.2.1(a) shows the coefficients and effect sizes from a model that compares children who experienced the intervention (RR or PI) with those who did not. The comparisons between the two groups of children are made both within schools and across schools. The important columns are those containing the effect sizes. Where a significant effect size is found, this suggests that the children who experienced the educational intervention (RR or PI) did significantly better than those that did not. The magnitude of the effect size is given in standard deviations. Thus a positive significant effect size of 0.4 suggests that those children who experienced the intervention, achieved 0.4 of a standard deviation better results than those children who did not. Table 4.2.1(a) demonstrates that the programme is effective for two sub-groups of children; those entitled to free school meals and those unable to read at six.

Table 4.2.1(a): Sub-sample of children taking FSM: reading and spelling skills at third follow-up (Autumn 1996)

Original sample	Reading/comprehension		Spelling	
	Coefficient	Effect size	Coefficient	Effect size
Reading Recovery				
Within-school comparison (n=45)	-7	-.10	0	Negligible
Between-school comparison (n=89)	30	.41*	.26	.24
Phonological Intervention				
Within-school comparison (n=56)	31	.38	.38	.30
Between-school comparison (n=79)	40	.48**	.49	.38*

Notes:

* = significant at 0.05 level; ** = significant at 0.01 level

Effect sizes were estimated from regressions controlling for initial scores on word reading test and diagnostic survey.

Within-school comparison: between RR or PI and control group in the same school.

Between-school comparison: between RR or PI and control groups in schools without RR or PI.

Source: QCA (1998) Table 4.10.

Cost data were collected by adding, to the RR and PI time per child, any additional help with reading provided to the child. The time data were converted into money costs, taking into account the actual salary costs of RR teachers and those giving PI. Otherwise, an annual teacher salary of £20,000 was assumed for 1992/3 and £21,000 for 1995/6 and 1996/7. The estimated costs are shown in Table 4.2.1(b) below. Obviously these costs must be combined with the estimated benefits of the interventions, to provide any guidance as to the cost-effectiveness of RR or PI, but they are interesting nonetheless.

Table 4.2.1(b): Estimated costs of RR and PI reading interventions compared to control group children

	1992/93*	1993/94	1994/95 (estimate)	1995/96 (estimate)	1996/97	1992-97 annual average
Reading Recovery schools:						
RR children	£890	£133	£173	£215	£244	£331
Control children	£120	£133	£133	£158	£158	£140
Phonological schools:						
PI children	£345	£227	£240	£272	£287	£276
Control children	£95	£267	N/A	N/A	£86	N/A
Control schools	£280	£293	£280	£315	£301	£294

Note: 1992/93 was the first year that the intervention took place.

Source: QCA (1998) Table 4.21

Costs are averaged over the five years in Table 4.2.1(b)⁶⁶. This indicates that RR was £37 more costly per year than normal provision (in the control schools), whereas PI was £18 cheaper than normal provision. However, staff training costs were not included⁶⁷.

The researchers attempted to compare LEA costs of SEN provision for children not being offered RR or PI, with the costs of RR and PI. Phonological intervention, followed by routine school provision, were the most cost effective options - producing significantly better spellers and marginally better readers at lower cost. For children who were still non-readers at six, RR was significantly better at marginally higher cost.

4.2.2 Interactive assessment and teaching reading programme

The second literacy study, by Nicolson *et al* (1999), was of a reading programme called Interactive Assessment and Teaching (IAT). This consisted of five key steps:

- an initial assessment of each child's reading in terms of meaning, phonics and fluency;
- on the basis of the above, priority teaching areas for the child were determined;
- a support plan was developed in steps for the target children;
- appropriate teaching methods were selected and implemented;
- each child's progress at each step was evaluated and recorded.

The study involved fewer children and a shorter time scale than the QCA study. IAT was implemented by two teacher-researchers recruited for the project, who used the IAT manual to design training for small groups of children in the four schools selected for the

⁶⁶ No present value calculations were made.

⁶⁷ The cost of training a RR teacher was about £1000 in 1995 and £300 for a PI teacher. The RR programme also requires maintenance and monitoring costs incurred by the RR national network and the LEA. The costs to each LEA involved of employing a RR tutor, professional development and management of the programme were estimated at £35,900 in 1995.

project. All children in Year 1 classes in the four schools were screened for reading performance, using WORD (Wechsler Objective Reading Dimension) reading and spelling tests. A trial group of sixty-four children were selected, aged between 5.5 and 6.5 years. These were the 16 children in each of four classes, in the four schools with the lowest reading performance, who also had a WORD reading standard score of 94 or less. A control group of 40 children was selected using the same criteria but from a different class in the same school, or from a different school of similar social composition if there were not enough children in the trial school. Control and experimental groups were matched for age and reading performance.

The target children worked in groups of 4 with a teacher, in two weekly sessions of 30 minutes each for 10 weeks. The main costs of intervention per pupil were:

- 0.5 hours teacher time per week for 10 weeks;
- 1 hour per child for applying reading and spelling tests;
- teacher planning time (not quantified);
- teacher time spent on ongoing assessment and record keeping.

The post-tests were taken after four months. Reading and spelling ability was measured by separate WORD tests. The results were reported as effect sizes (comparing mean differences of experimental and control groups) from two-way ANOVA, with F statistic for $p < 0.01$ given as significant. The effect sizes were measured in two ways: by reading/spelling score and by the age equivalent of these scores, as shown in Table 4.2.2. The authors state that the mean effect size of 1.71 for the experimental group, compared to 1.04 for the control group, is in line with that established in Reading Recovery.

Table 4.2.2: Mean effect sizes

Group	Reading standardised score	Spelling standardised score	Reading: age equivalent	Spelling: age equivalent
Experimental	0.94	0.95	1.71	1.24
Control	-0.53	0.24	1.04	0.48

Note: standardised score takes account of normal expected improvement (in chronological months) and so is less than non-standard score.

The authors claim a similar effect size (at immediate post-test) to Reading Recovery, for a programme the main cost of which is 3.5 hours extra teacher time per pupil compared with 35 hours per pupil for Reading Recovery. Hence the authors claim that IAT is much more cost effective than RR. However, they have not tested its effects over time.

4.2.3 Special educational needs (moderate learning difficulties)

Special Educational Needs expenditure has been estimated by the Audit Commission to account for around 10% of education spending. There is a substantial US literature on effective special needs provision for pupils with moderate learning difficulties but almost none for the UK. The DfEE commissioned a study, Crowther *et al* (1998), of the costs and outcomes for MLD pupils in special and mainstream schools. This reviewed available literature and also reported on a small study undertaken by the authors. From the literature review Crowther *et al* concluded the following.

Academic outcomes:

- these tend to be better in mainstream provision compared to special units/schools, but better mainstream outcomes depend on appropriate resourcing;
- full time placements in mainstream classrooms tend to be more effective than mainstream plus pull-out programmes;
- programmes which aim to provide structured interventions and support within the mainstream classroom appear to produce better outcomes than pull out with support outside the classroom.

Affective outcomes:

- academic self-concept is higher in special settings, but global self-esteem is less affected by type of placement;
- some evidence that mainstream students with moderate learning difficulties become more socially competent, though some experience rejection.

The empirical part of the study involved gathering data (between October 1997 and April 1998) on the costs and outcomes of SEN provision for MLD students in 33 schools (14 primary, 10 secondary and 9 special) in 8 LEAs. Sixty SEN professionals were involved in describing and categorising types of MLD student in order to produce six agreed categories. Schools and LEAs were then asked to provide information on the different resources they allocated to these 6 types of MLD. The study thus developed a useful costing methodology which may be replicated in future studies. The costs included were:

- costs of teaching and support staff in the classroom for the full week's timetable;
- staff costs of institutional support outside the classroom (SENCOs, pastoral support);
- external services (mainly from LEA);
- transport costs.

The data indicated very considerable variations in unit costs, even within mainstream schools, for pupils with the same type of MLD. For example, the cost of the least severe type A varied from £1,664 to £3,752 in mainstream primary schools and from £2,700 to £5,116 in mainstream secondary schools. This variation was partly due to differences in the funding of mainstream pupils in the different LEAs but was also, more interestingly, due to the different ways in which schools deployed teaching and learning support assistants (LSAs). This staff deployment accounted for two thirds of the cost of SEN provision. In the main, resources were allocated to students on the basis of perceived need or established practice, without any clear understanding of what outcomes could be expected from these resources. The study notes the growing deployment of LSAs without any research evidence on their effectiveness.

The study was unable to collect systematic data on outcomes. The researchers concluded that the allocation of resources to students with MLD was neither equitable (because resources did not consistently increase with severity of need) nor efficient, because schools and LEAs had no way of tracking the link between the resources deployed and outcomes achieved⁶⁸.

The study made a number of well-argued recommendations, in particular a shift in decision-making with respect to MLD students from need to outcomes. Resourcing should be tied to a specification of desirable outcomes and resource usage and subsequent outcomes should be systematically monitored. To assist this a national framework of outcomes for

⁶⁸ These findings are consistent with those of a smaller scale study of 8 schools in 2 LEAs by Marsh (2000). Marsh found that pupils with similar special educational needs, as measured by their reading age, received different amounts of resources even within the same LEA, and that some pupils with higher special needs received less additional expenditure than other pupils with a lower degree of special educational need.

MLD pupils should be established so that schools and LEAs can collect and analyse consistent outcome data and relate them to resource allocation data. The authors considered that much of the data needed already exist but need to be analysed at the appropriate level (individual, teaching group, cohort, school) and outcome and resource input data linked.

4.2.4 Summary

The findings of the small amount of UK experimental research on the cost-effectiveness of early literacy programmes, supplemented by the findings from the international research on specific literacy programmes, indicate that there are particular ways of deploying resources (*i.e.* specific resource mixes) which are effective compared with existing methods. However, the study by Crowther *et al* of special education resource allocation and special educational needs highlights the fact that there is a lack of systematic data and procedures by which researchers and practitioners can assess the cost-effectiveness of the thousands of interventions that go on daily in our schools.

4.3 OFSTED's qualitative assessment of schools' efficiency⁶⁹

OFSTED inspections make a qualitative judgement of a school using a common set of published criteria, which have been developed out of many years of HMI inspection. Consequently they are another important source of data and judgements about the efficiency and value for money of schools. In this section some of the evidence arising from this inspection process is summarised. The purpose of this section is to highlight another potential source of information on the issue of resource allocation and school performance. No attempt is made however to provide a systematic evaluation of the robustness or validity of the OFSTED approach. The focus is on the old OFSTED inspection framework that was in place prior to January 2000 rather than the new framework. This is because it is too early to see how the new framework will be applied in practice and as yet the new framework has only produced a small amount of data⁷⁰

4.3.1 Criteria for the efficiency of the school⁷¹

OFSTED defined efficiency in the standard way:

An efficient school makes good use of all its available resources to achieve the best possible educational outcomes for all its pupils – and in doing so provides excellent value for money. This section calls for a summative judgement on the basis of the findings on all aspects of the school's work.
(OFSTED, 1995b, p.121).

⁶⁹ The authors wish to thank OFSTED officials who spared the time to discuss the above issues with them and provided additional information on the evidence used by inspectors in making judgements of school efficiency. The account remains their own responsibility.

⁷⁰ The new framework differs in a number of respects from that in place prior to January 2000. In particular two new criteria have been added to the value for money judgement: the extent to which the school has improved, or not, since the last inspection; and the overall effectiveness of the school. However, it should be noted that discussions with officials indicate that the principles underlying the efficiency and value for money criteria outlined in this section remain largely unchanged in the new framework. The full details of the new framework can be found in OFSTED (1999a; 1999b).

⁷¹ The inspection schedule, prior to January 2000, actually covered four areas: management and efficiency of the school; education standards achieved by pupils at the school; quality of education provided; and the spiritual, moral, social and cultural development of pupils.

Given the absence of reliable quantitative measures of school output, the four criteria that inspectors were required to use for assessing the efficiency of a school were qualitative. These were:

- Criterion 1 educational developments are supported through careful financial planning;
- Criterion 2 effective use is made of staff, accommodation and learning resources;
- Criterion 3 efficient financial control and school administration;
- Criterion 4 the school provides value for money in terms of the educational standards achieved and the quality of education provided in relation to its context and income. (OFSTED, 1995a and OFSTED, 1995b, p.120).

Hence judgements were made about:

1. the aggregate relationship between output and inputs – family, peer group and school inputs (criterion 4);
2. the inappropriate mix of school resources (criterion 2 and criterion 1);
3. the quality of decision making – whether the school set appropriate priorities for the educational achievement of its pupils and allocated its available resources in the best way to meet these priorities (criterion 1);
4. the extent to which financial control and school administration were efficient (criterion 3).

Judgements were summarised on a scale of 1 (very favourable) to 7 (very unfavourable), with 4 being satisfactory. These numbers were not published in the inspection reports.

4.3.2 Judgements about the summary relationship between aggregate level of expenditure and school outcomes

Criterion 4, the value for money judgement, was reached by following a sequence of judgements, drawn from the full range of the inspection evidence for the school. The summary value for money judgement sequence is shown below in Table 4.3.2. Given its inclusion of contextual factors, outcomes, processes, and expenditure, the value for money judgement was, in effect, a qualitative interpretation of an education production function.

Table 4.3.2: Criteria for reaching summary value for money judgement

CONTEXTUAL FACTORS	JUDGEMENT RECORDING GRADE						
	1	2	3	4	5	6	7
Socio-economic circumstances of pupils							
Attainment on entry							
OUTCOMES							
Pupils' attainment in relation to national averages or expectations							
Pupils' progress							
Pupils' attitudes, behaviour and personal development							
PROVISION							
Quality of education, particularly teaching							
EXPENDITURE							
Unit cost for the type of school							
VALUE FOR MONEY JUDGEMENT							

Source: (OFSTED, 1995a) and OFSTED (1995b, p.125)

The value for money judgement was based on a combination of school level and external national data. The type of data used and the balance between school level inspection data and external data differed between criteria.

Contextual factors were socio-economic variables taken from the population census of wards near the school and data on the proportion of pupils known to be eligible for free school meals. Attainment of pupils on entry was derived from LEA and school data.

Pupil outcome data were largely derived from the key judgement recording statements produced by the inspection and raw test and examination results. QCA benchmarking data (published in PANDAs) was used, which compared the school's at KS1 to KS4 results, with the inter-quartile ranges of 'similar' schools (*i.e.* those falling within the same range of FSM eligibility). Schools' own value added data on pupil progress, when available, were also used.

The quality of provision was judged using the inspection evidence with particular emphasis on the quality of teaching.

Expenditure per pupil, was calculated from the information given on the Headteacher's Form as the total expenditure from all sources divided by the pupil roll. Inspectors were given further information about the proportion of local schools budget delegated by LEAs and allowance was made for additional GM/foundation school funding. Unit costs were then compared to the inter-quartile ranges for all schools that completed Headteacher Forms in last full year of available data. Unit costs were not differentiated by size of school, or degree of social disadvantage, though these are both factors associated with higher unit cost⁷².

⁷² According to verbal information supplied by OFSTED, a potential bias towards poorer value for money judgements for small schools, and for schools with high proportions of socially disadvantaged pupils, was offset by inspectors using their judgement in relation to criteria 1 and 2 (above). Inspectors assessed how well the school used available resources to fund its educational priorities and how appropriate these priorities were in relation to inspectors' diagnosis of the school's strengths and weaknesses.

4.3.3 Judgements about resource mix within the school and the quality of decision making

Criteria 1 and 2 were particularly important for assessing the efficiency of resource allocation within a school⁷³, which was judged in terms of the quality of decision-making about resource use and of the deployment of resources. Inspectors made use of external data to provide performance indicators of key cost ratios (*e.g.* percentage of budget spent on teachers, support staff, administrative staff, learning resources).

The Inspection schedule stated that:

Inspectors should examine the use of the funding for different purposes, including provision for teaching staff, curriculum development, learning resources and premises. Some comparison of costs per pupil and the percentage of income spent on different items can be derived from the PICS report, which gives information on the range and median levels of expenditure. These need to be interpreted in the school context. Comparison must be tentative because of differences between LMS schemes throughout the country and because of differences in other income. The circumstances of schools also differ, particularly in relation to teaching staff and premises costs.

(OFSTED, 1995a and OFSTED, 1995b, p.122.)

Criterion 2, the effective use of staff, accommodation and learning resources, was assessed in relation to:

- the number, qualifications and experience of teachers and class room staff relative to the demands of the curriculum;
- induction, appraisal and INSET;
- the adequacy of the accommodation for the number and ages of pupils and for the required range of curriculum activities;
- the extent to which learning resources were appropriate in range, quality and quantity for the curriculum and range of pupils and how well these resources were deployed.

Particular importance was given to criterion 1 – namely whether the school could demonstrate to inspectors that its educational priorities were well chosen in relation to the strengths and weaknesses of the school as assessed by the inspectors, that its resource allocation was rationally planned in relation to addressing these priorities and that plans were well implemented, monitored and reviewed. A school's efficiency was not judged in relation to a common set of cost ratios or input-output measures, *i.e.* it was contingent on the specific circumstances of each school.

4.3.4 Assessment of efficiency and value for money achieved by schools nationally

OFSTED's Annual Reports on the quality of schools' educational provision included a statistical summary of the findings concerning school efficiency, reproduced in Table 4.3.4. This shows that 50% of primary schools, compared with 57% of secondary schools, were judged to be providing good value for money in 1998/99 – an increase compared with 1996/97. 8% of both primary and secondary schools were judged to be providing poor value

⁷³ Criterion 4 was less important since well-administered procedures for financial recording and reporting, though necessary for providing good management information for resource allocation, do not of themselves guarantee good quality decision-making.

for money in 1998/99. There had been no improvement in this proportion over the previous two years.

The application of the OFSTED framework provided an annually updated summary assessment of the distribution of schools in terms of their relative internal efficiency in allocating resources to produce educational outputs. Given the nature of the criteria used and the data available, OFSTED inspection data could not provide evidence of the impact of differences in the amount of expenditure per pupil (and hence in the quantity of resources per pupil) on the educational outputs of schools.

Table 4.3.4: The efficiency of schools in England 1998/99 and 1997/98 (in parentheses)

	Good %	Satisfactory %	Unsatisfactory %
PRIMARY SCHOOLS			
Efficiency ¹	63 (52)	33 (41)	4 (7)
Financial planning (criterion 1)	60 (50)	30 (34)	10 (16)
Use of teaching and support staff (criterion 2)	59 (51)	36 (41)	6 (9)
Use of learning resources and accommodation (criterion 2)	57 (51)	40 (45)	3 (5)
Efficiency of financial control & administration (criterion 3)	78 (71)	20 (25)	2 (4)
Value for money (criterion 4)	50 (39)	43 (52)	8 (8)
SECONDARY SCHOOLS			
Efficiency ¹	65 (61)	29 (32)	6 (7)
Financial planning (criterion 1)	62 (59)	26 (25)	11 (16)
Use of teaching and support staff (criterion 2)	57 (55)	33 (35)	10 (10)
Use of learning resources and accommodation (criterion 2)	60 (58)	37 (36)	3 (6)
Efficiency of financial control & administration (criterion 3)	85 (82)	13 (15)	2 (3)
Value for money (criterion 4)	57 (52)	35 (41)	8 (7)

Note 1: Efficiency is a composite rating for inspection schedule 6.3, the efficiency of the school, derived from ratings for criteria 1 to 4.

Source: OFSTED (2000) Annex 4 and OFSTED (1999) Appendix 3, pages 74 and 79.

4.4 UK education production function and cost-effectiveness research: conclusions

The UK research literature illustrates well the important points made in Section 1 that theory, model specification, data, statistical method and the quality of the resulting empirical evidence are intertwined.

The evidence in favour of a positive impact of resources on educational outcomes from UK education production function studies can be summarised quite briefly. The overall level of spending per student was found to be significantly and positively related to educational outcomes in only one of the studies reviewed – for female wages (Dearden *et al*) Indicators of spending on teachers have fared better. Secondary teachers' salaries per pupil were significant and positive for male wages at 23 (Dearden *et al*). The pupil-teacher ratio was found significant and correctly signed for staying on at school (Dustmann *et al*), and for

exam results by Dolton and Vignoles and Dearden *et al*⁷⁴. There is almost no UK evidence from quantitative research that smaller class size leads to better outcomes. However, a proxy for teacher quality was found significant by Lord (proportion of graduate and equivalent teachers) and by Bradley and Taylor (proportion of part-time staff). There is some limited evidence that school size is positively related to exam performance (*e.g.* Bradley and Taylor for GCSE and O'Donoghue *et al* for A-level).

School type appears significant in explaining examination performance in all the NCDS studies, as well as in Bradley and Taylor, and O'Donoghue *et al*. However, it is not clear to what extent this is due to the peer group effect, to better resourcing or better teaching quality in grammar, independent and single sex schools.

The evidence on the cost-effectiveness of A-level provision also indicates considerable differences by institution, largely due to differences in group size. Whether the higher cost of school sixth forms compared to the FE colleges is compensated for by greater value-added is still not clear because of data problems in reaching an equivalent cost-effectiveness measure.

Studies using the NCDS dataset have also begun to hint at the importance of process variables and the differential impact of resources on different types of student (*e.g.* the significance of peer group effects in Feinstein and Symons and O'Donoghue, and of interaction effects in Dearden *et al*). Findings such as these suggest that over-simplistic model specification may conceal the impact of resources on outcomes or fail to allow sufficiently for the complexities of classroom interactions on student outcomes.

The limited evidence on early literacy interventions suggests that particular ways of utilising resources are more cost-effective than others. The study of the effects of different resource mixes in schools is almost uncharted territory in UK research. Crowther *et al*'s study is useful in highlighting the absence of systematic ways of evaluating different resource mixes, which result in inefficient resource allocation practices. Schools and LEAs are collecting and recording data on student outcomes and on resource use but these are not brought together on a consistent basis to enable meaningful analysis to take place. This is strongly reiterated in Mayston and Jesson (1999).

An alternative approach is the qualitative framework adopted by OFSTED. This allows structured judgements to be made on how well schools use their resources. It has highlighted differential efficiency between schools. For example, the OFSTED inspections data base shows that schools are judged as being differentially efficient (see Table 4.3.4). OFSTED's study of effective sixth forms (OFSTED, 1996) and the Audit Commission's 1993 study also provide evidence of differential efficiency. So do Crowther *et al* who conclude that schools' resource allocation practices with respect to SEN are not efficient. School effectiveness research provides additional indicative evidence.⁷⁵ If schools are differentially effective in producing educational outcomes and are similarly resourced this implies differential efficiency. This assumption is realistic for schools from the same LEA, which applies to some of the UK school effectiveness studies.

In summary, despite the conviction of educators and most parents that more resources are required to achieve higher educational outcomes, and hence that unequal resource distribution between schools with equivalent needs is inequitable, UK research has failed to come up with unequivocal support for this belief.

⁷⁴ For men attending secondary modern schools, for lower ability women and for women's wages at 33.

⁷⁵ UK research has shown that between 8 to 15% of the variation in pupil outcomes is due to between school differences, after controlling for pupil level factors (Teddlie and Reynolds, 2000).

5. The Ideal Research Project

This Section draws on the previous arguments presented in the review to provide guidance to policy-makers about a future high quality programme of research in this field. The section begins by highlighting some of the major theoretical and empirical difficulties in this field, as outlined earlier in the report. This gives some motivation for an ‘ideal’ research project. Section 5.2 then discusses specific methodological considerations before focusing on the key variables required and highlighting the main data issues. Finally the Section ends by discussing how the VFM Unit’s GM Schools Database⁷⁶ and other DfEE databases might be suitable sources of data for an ideal research project. Whilst the main limitations of existing data are highlighted, a comprehensive and systematic evaluation of the quality of each data set is not provided as this is outside the remit of this review. Furthermore, time and space constraints do not permit a fully specified research design to be drawn up, however the main requirements of a high quality but realistic study of the effect of school resourcing are sketched out.

5.1 The need for an ideal research project

As has been discussed, the specification of the education production function in the existing literature tends to be of the ‘black-box’ variety, whose nature in part reflects the absence of well-established theories of how changes in resources (*per se*) impact on school processes and, through these, affect children’s learning. In other words, education production function research is not based on well-specified technologies for teaching and learning.

Furthermore, current research often fails to acknowledge that the amount of expenditure will not have a determinate impact on student outcomes if the *quality* of resources a given amount of expenditure purchases is variable between schools, and if the mix of resources used in schools is varied. An additional complication is that the teacher-pupil and pupil-pupil interactions in the classroom that affect learning are dependent on personal and organisational characteristics, which may bear little relation to the physical quantity of resources deployed. Hence specifications of the education production function which only link the quantity of inputs to the quantity of outputs will omit important intervening variables.

Theoretical foundations should suggest not only what variables should be included in an estimated model, but also the mathematical formulation of relationships between the variables. Perhaps it is the lack of theoretical foundations in this field that has led to almost all research assuming that the relationship between inputs and outputs is linear or log linear, that it remains proportional as resources increase, and that resources have the same impact on all students and contexts. These are assumptions that need more rigorous testing.

The UK literature specifically also has a number of additional problems. Many studies have used data aggregated at the LEA or school level, and are therefore vulnerable to aggregation bias. Studies that have used student level data all draw from the NCDS. Consequently, they are focused on the individual rather than on the individual nested in a school, as in school effectiveness studies, and have few school level variables (due to data limitations). Much of the more recent UK research has acknowledged that the basic model is a simultaneous equations model, with the resources which a pupil receives in part dependent on his/her family background and prior attainment. However, only relatively few studies have actually attempted to tackle this endogeneity problem.

⁷⁶ Formerly the Funding Agency for Schools (FAS) database.

5.2 Methodological considerations

The methodological issues discussed above and in Section 1 therefore suggest the following broad guidelines for future research in this field.

- 1) Future empirical research should be more closely linked to educational theory. Educational theories, such as those of Carroll (1963) or others outlined in Teddlie and Reynolds (2000), can help researchers to identify more clearly the ways in which students' learning outcomes might be affected by how the various inputs enter into the educational process, and indeed by the way in which these inputs are processed through the system. In particular, education production function models need to relate to theories of school organisation, teaching and learning. A better understanding of how resource levels and mixes relate to the creation of effective learning environments is needed. This requires interdisciplinary work between economists and educational researchers, in particular those researching teaching and learning and school effectiveness. Future empirical work needs to rigorously test the implications of these various educational theories. In fact simply identifying clear and testable hypotheses from educational theory will tend to provide greater coherence to future research in this field. Greater use of educational theory will also help draw together strands from two generally separate literatures, namely the School Effectiveness research field (Teddlie and Reynolds) and the Educational Production Function literature that has been reviewed in this report. As was pointed out in Section 1, the more effective use of educational theory should also reduce the probability of any new research project simply repeating previous empirical work.
- 2) There is a clear need for more methodological work in this field, both in terms of improving and combining existing techniques and developing new evaluation methodologies. For example, more work is needed to further develop the application of statistical techniques needed to overcome the potential endogeneity problem in this literature (also see point 5 below). Equally, data collection methods in this field need to be determined by systematic evaluation methodologies rather than be conducted on an *ad hoc* basis (Section 4.3).
- 3) Future studies need to more rigorously assess the empirical implications of using different techniques, such as OLS regression versus DEA. This would involve comparing the results from different techniques and thoroughly testing the assumptions required by the different methodologies. Although it has been noted that DEA and regression techniques for example, serve somewhat different purposes and may therefore be complementary. The important point is that both techniques need to be compared, within one study, using the same data and testing the same hypotheses. This means that a future research project should obtain pupil level and school level data, and use both standard regression analysis (on pupil level data) and DEA (on the school level data), and compare the results from the two approaches. Only when further systematic research has been carried out, using a range of techniques, will this question be fully answered.
- 4) The most major specific methodological problem is the potential endogeneity of school resourcing levels. Future research could take a three-pronged approach to tackling this issue. First, researchers need to identify potential instrumental variables. In addition to using instrumental variable estimation within a conventional regression framework to overcome the endogeneity problem, IV methods might also be combined with other techniques, such as multi-level modelling.⁷⁷ Another possible way to overcome the

⁷⁷ There appears to be no research that has combined IV methods with a multi-level modelling approach.

endogeneity problem is to estimate structural simultaneous equation models. In the UK, these models might use factors, such as government spending criteria (SSAs and LEA funding formulae), that influence expenditure per pupil, for example, but do not directly affect learning outcomes. Finally, policy-makers need to consider the use of randomised experiments, particularly in the context of evaluating specific educational interventions such as Literacy Programmes.

5.3 Key variables required for a high quality research programme

It has already been argued that an ideal research project should use student level data, with a value-added formulation. The research should be based on testable hypotheses, rooted in educational theory, and researchers need to rigorously test the sensitivity of results to different methodological approaches. However, these are quite general guidelines. This section identifies the important variables required to carry out such a research project, starting with outputs, then focusing on resource inputs and finishing with other important explanatory variables. In order to determine the necessary variables, some of the main research questions that have yet to be fully answered in this literature, particularly in a UK context are highlighted.

Main research questions

1. Does the level of resourcing (expenditure per pupil) affect pupil outcomes?
2. Does the mix of resources affect pupil outcomes? In particular, are there combinations of resources that are more efficient than others?
Subsidiary questions: Does the use of LSAs improve pupil outcomes? Can LSA substitute for teacher time?
What is the trade-off between non-contact time for teachers and larger class size?
What effect does investment in ICT have on pupil outcomes? Can ICT substitute for teacher time? Can ICT improve teacher productivity?
3. How does the quantity and mix of resources affect educational outcomes for pupils with special educational needs?
4. How do pupil variables interact with resource variables and affect pupil outcomes?
Subsidiary questions: what is the importance of compositional/peer group effects on pupil outcomes?
Do resources impact differentially on pupils with different characteristics?
5. Are schools differentially efficient and to what extent?
6. What distinguishes resource management in more efficient from less efficient schools?

5.3.1 Output variables

This section considers the main output variables of interest in relation to questions asked above. Most of the studies that have been reviewed have used standardised test results, years of schooling or examination results as outputs. The value of these indicators is that they are factors which educators can reasonably be expected to have an influence over. They are the outcomes that policy-makers and parents alike focus most on, and hence can be considered legitimate output variables in any analysis of teacher or school performance. Furthermore,

since the use of these outputs is common in the literature, this allows researchers to compare their results with previous work in the field.

Future work in this field should therefore continue to focus on these primary output variables, although where educational theory dictates, other output variables might also be considered. For example, since various educational theories suggest that time on task is important for learning, other outputs of interest would include truancy rates or behaviour in the classroom. In general we conclude that cognitive, affective and life chance outcomes (destinations) should be measured, as well as intermediate outputs such as attendance and exclusion. Measures of cognitive outputs are available from national tests⁷⁸ and examinations and are available from existing data sets. In this respect the UK is fortunate in having national examinations and tests at different stages of schooling which provide reasonably consistent national data on cognitive outputs in the major domains. These have better construct validity for what is taught in schools than standardised cognitive tests. However, there are problems such as devising a common scale for A/AS and vocational qualifications (DfEE, 1998b) which have, for example, impeded reaching conclusions about the cost-effectiveness of different forms of post-16 provision. Affective outcomes would need special instruments to be administered in a sample of schools.

A number of important caveats apply here. First, caution needs to be applied when using newly developed test and examination results that have not yet been validated, in terms of standards. For example, the newly introduced 16-19 qualifications need to be validated by comparing results over time and referencing to the standards of other types of examinations. The second caveat is that this type of research must always be used cautiously for policy purposes. If a particular outcome (Grades A-C at GCSE) becomes the main policy focus, it is well known that this may skew the incentives of teachers and schools towards those who, with a bit of effort, might meet this standard, away from those at the extreme top and bottom of the academic distribution. For this reason sensitive application of targets or output objectives is crucial.

Finally, it should be remembered that the direct economic effects of education are generated via their impact on labour market outcomes. Focusing purely on the relationship between education inputs and pupil outcomes therefore provides only a partial picture of the efficiency of the school system. More work needs to be done on the impact of school resourcing on labour market outcomes if a fuller picture of school efficiency is to be obtained.

5.3.2 School resource input variables

As has been emphasised, there is an extremely limited body of high quality UK evidence to draw inferences on what are the most valuable measurable resource input variables in the education process.

While raising **expenditure per pupil** may be beneficial, in the absence of methodologically sound evidence, it is not possible to make a definitive claim and therefore further work investigating this input is needed. In particular an ideal study needs information on total funding per pupil per year (from all sources), as well as disaggregated expenditure on teaching staff, classroom support staff, administration (staff, services, non-staff items), costs of providing physical environment (maintenance, utilities *etc*), educational resources, professional development, value of centrally retained services and annualised value of capital assets.

The evidence suggests that reducing **class sizes** does not in itself appear to be a cost-effective means of raising student outcomes. However, investigating how smaller class sizes

⁷⁸ Not all cognitive outcomes may be measured by national tests *e.g.* ICT skills and numeracy and literacy skills of some students with MLD.

and other real resources, in a variety of different settings, might or might not improve outcomes is still a useful avenue for future research. For example, the use of smaller classes for less and more able students or the use of smaller classes with different types of teaching techniques (whole class versus group work) could be investigated. Specific variables of interest here would be those measuring the real resources used at school level: pupil teacher ratio, average class size, non-contact time of teachers, management time, length of school day, support staff hours, administrative staff hours, computers, books, space for specialist provision.

Investigating **teacher inputs (teacher characteristics and behaviour)** provides one of the most promising avenues for future research. This research will also have important implications for resource use in schools (*e.g.* are more experienced/educated and hence more costly teachers substantially more effective). It will also have implications for the cost effectiveness of any particular policy designed to raise teacher and school performance (*e.g.* performance related pay). Key teacher inputs of great interest would be qualifications, experience, posts of responsibility, professional development record, cognitive ability test scores and assessment of teaching effectiveness.

Lastly, in the UK context, further information on LEA inputs would be needed. For example, data on the external services provided via LEA retained expenditure, the overall level of funding at LEA-level and details of the formula for devolving individual school funds. Information on the Local Schools Budget retained and its allocation (*e.g.* special educational needs) would also be useful.

In general the research reviewed here suggests that richer data will be able to give a more illuminating view of the importance of various school inputs. Hence an ideal study would need to consider the following.

- The use of more refined measures of inputs, such as specifically examining initial years of teacher experience, and looking at teacher qualifications by subject area, rather than simply examining the total number of years a teacher has spent in the profession.
- Greater investigation of the impact of measurement error is required, particularly as it pertains to the measurement of expenditure information.
- Other factors that have been rarely employed in the literature, such as teacher behaviour and techniques, may also provide useful information (see Section 2.2.3 and Goldhaber and Brewer, 1997).

5.3.3 Explanatory variables

A 'standard' list of explanatory variables would include:

- **pupil characteristics** (gender, age, ethnicity, family background variables⁷⁹ to allow for the quantity and quality of parental inputs – particularly parental education and any proxies for the time that the parent spends with the child on learning activities, pupil achievement on entry into school⁸⁰ or at the start of particular educational intervention⁸¹).

⁷⁹ Most studies use free school meals entitlement but other background variables would ideally be needed such as parental interest, number of siblings and expenditure at home on educational resources.

⁸⁰ Existing data that can be used include: base line tests on entry to school (but there are 90 or so tests approved by QCA); KS1 and KS2. KS3 GCSE, A/A-level, GNVQ, NVQ. One might also use common standardised cognitive tests such as those used by NFER and YELLIS/MIDYS and PIPS (University of Durham). These have advantages over KS tests in being more differentiated and standardised. Their relationship to future pupil attainment has been much more thoroughly tested than SATs. However, they are not used by all schools.

⁸¹ Very few studies have obtained this kind of information. Mother working, and information on parental help with homework have been used as proxies instead.

- **neighbourhood variables** (socio-economic profile of the local area to allow for peer effects).
- **school variables** that are not directly related to resource inputs (school size, degree of selectivity, socio-economic profile of students).

Such variables are typically viewed as controls from a policy perspective since they are factors that are not directly influenced by professional educators and therefore cannot be used as policy instruments. As has been emphasised in the literature however, such factors are clearly important, and the omission of such variables is likely to cause spurious results.

Pupil achievement at the beginning of the educational intervention is perhaps the most important of the control variables for value-added models. In the UK, the absence of a national common set of prior attainment measures has restricted education (and school effectiveness) research. National data sets linking prior attainment to outcomes at student level exist so far only for GCSE to A-level, or Key Stage 3 to GCSE. (The latter has the disadvantage of measuring progress within the school from a baseline achieved within the school.) School effectiveness research at Key Stage 1 to GCSE has been done only for LEA datasets or by commercially provided value added services to schools (for example, University of Durham). The paucity of accessible student level achievement and prior attainment data sets has forced researchers to use school level examination data with the proportion of students eligible for free school meals as a proxy for prior attainment, with the attendant omitted variables and aggregation bias problems.

5.3.4 Interactions

It is also important to reiterate that there may be interdependencies between the explanatory variables and school input/resource variables, and indeed between the different resource input variables. For example, more effective teachers may set homework that also raises the value of time spent by parents on home education. Another important question involving the interaction of variables is; whether students from an under-privileged background gain more from higher educational inputs than more privileged students. Viewing the explanatory variables listed in the previous section as mere controls may seriously underestimate the complexity of the education process and a thorough examination of the important possible interactions suggested by educational theories may yield valuable policy inferences. A profitable avenue for research would therefore entail a shift away from examining individual inputs in isolation and looking at how best to combine resources effectively. While previous researchers have sometimes explicitly or implicitly emphasised the importance of the interactions of resource inputs, far fewer have actually tested systematically for such interactions (with the notable exceptions of Wright, Horn and Sanders, 1997 and Dearden *et al*, 1997).

A thorough investigation of possible interactions would entail:

1. Testing the commonly held, but largely untested propositions, that the interaction of different resource inputs potentially provides larger gains. For example asking questions like: does putting a more experienced and motivated teacher in a larger class lead to an equivalent student outcome to placing a less experienced and motivated teacher in a smaller class? Do the differing characteristics of students *between* and *within* schools, such as a higher proportion of lower achieving students, mean that hiring more qualified teachers leads to better student outcomes?

2. Can researchers determine ‘optimal’ mixes of inputs to maximise educational gains. For example, is there an optimal mix of teacher experience, class sizes, and expenditure for schools in the UK?

In addition, the bulk of the work reviewed concentrates exclusively on the *benefits* of certain inputs. Most studies make little attempt to examine the cost of interventions with respect to input variables. The lack of such cost-benefit analysis is a severe limitation in the literature.

5.4 Data issues

A recurring theme throughout this review has been the need for high quality, student level, data. This problem has certainly undermined some of the existing evidence on the effects of school resourcing, particularly in the UK. Whilst there is no need to repeat the comments made in Mayston and Jesson (1999), regarding the need for a National Pupil Database it is useful to re-iterate the key data problems in this field and these issues certainly strengthen the case they made for the NPD.

There are a number of key data issues.

- 1) Many studies suffer from bias due to the aggregated nature of their data and because they do not use a value-added formulation. An ideal research project therefore needs student level data, in addition to more aggregated measures of school resourcing, as well as information on students’ prior attainment.
- 2) In the UK good use has been made of longitudinal studies, like the NCDS, which contain data on students’ prior achievement and student level data on school resource inputs. However, since these longitudinal surveys were not designed to specifically look at the effect of educational interventions, they lack the necessary detail. Hence, the surveys can only be used to look at a few key inputs (pupil-teacher ratio), ignoring other inputs relating to the school environment that may also be important, and leading to omitted variable bias.
- 3) At the other end of the spectrum, the administrative databases (such as Form 7 and OFSTED data) generally provide very detailed information about school inputs, both qualitative and quantitative. Yet these data sets are not linked to student level data, so researchers are unable to obtain information on students’ prior achievement levels and other background characteristics. Without such links, UK research in this field will be limited by the need for researchers to obtain primary data themselves, a costly process. This is the reason that data collection and evaluation issues have already been given the highest priority by the Centre for the Economics of Education (Vignoles, Desai, and Montado, 2000).
- 4) School resource data for education production function research are currently very limited despite a decade of local management of schools. The only national sources utilised by education production function studies are the Annual School Census (Form 7) and LEA returns (CIPFA’s Education Statistics: Estimates and Actuals). Otherwise researchers have had to gather data through fieldwork, which limits sample size (*e.g.* Thomas, 1990; Audit Commission, 1993). The lack of a usable dataset of expenditures and resources at school level has severely restricted the range of resource variables, leading to bias caused by omitted variables, measurement errors and aggregation.

5.4.1 Current data sources

As it stands, official DfEE sources of data, including the VFM database⁸², need some further development in order to meet the demanding data requirements of the 'ideal' high quality research programme outlined in this section. This is because no one data set contains the following basic data required to carry out such a research programme; a) student level achievement records with baseline achievement levels, b) basic information on each pupil's socio-economic background, c) data on each pupil's school (*e.g.* school size, gender mix, average funding level, socio-economic profile of students) and d) the resources applied to that particular pupil (*e.g.* their mathematics class size, or the education level of their English teacher). Furthermore, at the moment it is not possible to merge different data sets in order to create such a data source.

More positively however, there are a number of sources of very rich high quality data that are available. For example, the VFM database provides very comprehensive school level data. This database is particularly useful because it contains both Form 7 information and some school level performance data. Since it also contains the school name, DfEE number and LEA number, these data could be easily merged with additional data, such as that held by OFSTED. Nonetheless, the VFM database, like others held by the DfEE, is crucially limited for the purposes of research by the fact that it does not contain any pupil level background information.

The development of National Pupil Database that will link pupil level GSE/SAT scores with pupil level annual School Census data, will be a major advance for researchers. This is because it will allow pupils' performance over time to be linked to pupil/school context variables available from the ASC. However, the NPD will not overcome the problem faced by researchers of poor data on school resourcing. Pupil level data from the NPD will still need to be linked to school resourcing data if full scale research into the relationship between school inputs and pupil outcomes is to be carried out. The closure of the Funding Agency for Schools means that detailed resourcing data at school level are now only collected by OFSTED. Although it should be relatively easy to merge these data with the NPD, OFSTED only inspect 25% of schools within a single year limiting the generalisability of any findings. Consequently it is essential to do more to develop a financial reporting framework for all schools which will provide consistent, comprehensive and detailed breakdowns of school expenditure if a fully rigorous research programme is to be carried out.

Although discussions are still ongoing it looks as though in its initial phase the NPD will also only provide limited information on teachers. This is because a) current information from the DfEE suggests that it will not be possible to identify each individual pupil's teachers from the records in the NPD and b) even if it is possible to link pupils and teachers, the NPD will not contain sufficient detail about each teacher to enable researchers to investigate issues such as the impact of teachers' education levels on pupil outcomes. Consequently this will limit the questions that can be explored even if suitable financial data can be combined with the NPD.

There are, however, ways in which the teacher quality data can be improved in the short-term which do not rely on expanding the scope of the NPD. In particular the Database on Teachers Records contains a wide range of quality measures such as experience, class and type of degree for individual teachers. These records also have a school identifier. It should therefore be possible to widen the range of teacher quality variables available at the school level. It is clear though that this merging would be resource intensive, given the current

⁸² This is a comprehensive database holding information on GM schools Income and Expenditure accounts by detailed category breakdowns, as well as Key Stage 3 and GCSE performance data for all non-independent schools in England, and School Contextual data from Form 7. Data are held from the 1993/4 financial year onwards.

format of the DTR. Updating the Schools Staffing and Curriculum Survey and merging it with the NPD data would perhaps be an easier way to proceed. One advantage of using the SCSS would be that extra questions could be added to the survey to get exact measures of the key variables in which researchers are interested. However it should be noted that only a limited number of schools are included in the SSCS.

The NPD might also be linked to other data sets to provide additional information on the family background of pupils, and on their neighbourhood and local environment. For example, NPD records might be merged with data such as NOMIS or Census information on the basis of school postcodes (or ideally pupil postcodes if available). However, obtaining information on pupil's family background is likely to prove more difficult unless basic information such as gender, age, ethnicity and parental occupation or education level is included in individual student records.

5.4.2 Ways forward in the short term

The developments outlined in Section 5.4.1 are likely to take a considerable period of time to come to fruition.

In the shorter term there are a number of possible ways forward.

- 1) First the current piloting of the pupil-level Annual School Census could be used to obtain more comprehensive information from a subset of schools. These data could then be linked to pupil-level GCSE/SAT data currently collected for the Autumn Package. Where financial data are available these could be linked in too.
- 2) Alternatively, a high quality longitudinal survey of a sample of pupils, teachers and schools, representing as wide a range as possible of per pupil expenditure for given SES levels could be constructed. This survey would need to be carried out over a period of 2-3 years. Such a survey would of course be relatively resource intensive, but would certainly be the method most likely to generate all the necessary information within a short time period. The survey could also be supplemented with data from various secondary sources, including OFSTED, in order to minimise the size of the survey and its cost. This research design, since it would involve data collection from individual schools, could also help to pilot further developments in software applications for more efficient data collection and reporting in schools.
- 3) A final alternative would be an experimental research project. An experimental research design would go a long way to solve the endogeneity problem, by assigning schools to two random groups and introducing changes for one group and not the other. An experimental design is particularly well suited to finding out whether specific educational interventions will have an impact in this way on pupil outcomes. For example, it could be applied to investigate the impact of learning support assistants within an education production framework. This could explore not only whether employing more LSAs is effective and efficient, but also the most cost-effective ways of using and training LSAs. It should be noted that such an experimental study would still need to gather some data on pupil characteristics and school and class level variables, but in a trial and control setting. Another, more controversial, possibility is to fund some schools an additional amount and not others and investigate whether there are differences in pupil outcomes as a consequence.

Finally it should be emphasised that, if the findings of a research programme are to have the maximum impact on improving the efficiency of resource allocation and deployment

in schools and LEAs, the evidence has to inform practice. This means that practitioners must engage with the evidence, find it convincing and be willing to let it inform their decision-making. Research designs which involve practitioners (such as suggested in point 2 above) have a much better chance of producing richer and more convincing findings and of influencing practitioners in the choices that they make.

6. Conclusion and Implications for Policy

This review started with an assessment of the methodological difficulties involved in measuring school efficiency and concluded that:

- a) future empirical research should be more closely linked to educational theory;
- b) there is a clear need for more methodological work in this field, both in terms of improving and combining existing techniques and developing new evaluation methodologies;
- c) data quality is a pressing issue, particularly in the UK;
- d) there needs to be more work on comparing different techniques, such as OLS regression versus DEA and other methodological approaches;
- e) the most major methodological problem that still needs to be overcome is that of the potential endogeneity of school resourcing levels. The use of IV methods, simultaneous equation modelling and random experiments needs to be considered further in this context to overcome the endogeneity issue.

With these methodological difficulties in mind, recent 'high quality' international research in this area has been reviewed. Taken as a whole, the international literature suggests that some measurable school inputs do matter: potentially class size, teacher experience and teacher salaries. However, the magnitude of the reported associations has been quite small. The evidence on specific educational interventions is more optimistic; most schemes considered generated substantially improved student performance.

The UK literature review showed that from 8 to 15% of the variation in pupil outcomes is due to between school differences, after controlling for pupil level factors (Teddlie and Reynolds (2000)). The overall level of spending per student was found to be significantly and positively related to educational outcomes in only one of the UK studies reviewed. The overall pupil-teacher ratio of schools was found to be significant and correctly signed in several studies. However, there is almost no UK evidence that smaller class size leads to better outcomes. Although school type appears important in explaining examination performance; it is not clear to what extent this is due to; the peer group effect, to better resourcing, or better teaching quality in the different types of schools. The UK literature review also hinted that ignoring interactions between resources and other inputs may conceal the impact of resources on outcomes or fail to allow sufficiently for the complexities of classroom interactions on student outcomes.

Finally, guidance has been included in this review for future research into School efficiency. The most important conclusion was that the available data is insufficient to carry out a high quality study that would overcome most of the methodological problems identified in this review. Once the National Pupil Database comes fully on line, it will meet a number of the methodological criteria that have been identified. A separate exercise would still need to be implemented to ensure that sound and consistent financial data are available. The National Pupil Database also needs to have the facility to link effectively into the available teacher databases and other sources of information on pupils' neighbourhoods and local environments. The need for the National Pupil Database to eventually contain important

basic background information on each pupil (gender, age, ethnicity, and parental education/social class) has also been highlighted.

7. Glossary of Key Terms⁸³

Correlation – a statistical measure of the closeness of the relationship between two variables. A high correlation suggests a very close relationship between the variations in the value of one variable and the variations in the values of the other.

Data envelopment analysis – sometimes called Farrell frontier methodology. It is a non-parametric technique used to estimate the production frontier, *i.e.* to estimate the relationship between one or more inputs and one or more outputs.

Dirty data – see **measurement error**.

Educational production function or frontier – the function or frontier is the mathematical relationship between the output of a school or education system and the inputs or factors of production used to produce that output. It can be interpreted as a technical relationship, which describes the set of efficient transformations between inputs and outputs, for a given technology.

Endogenous variable/endogeneity – a variable whose value is determined by the other variables within a system. For example, school quality may be endogenous if it is determined by other variables (*e.g.* family background) in the system.

Error term – sometimes called the disturbance term. It is a random (stochastic) variable that has well-defined probabilistic properties. It represents those factors that affect the dependent variable but are not or cannot be taken into account by the independent variables.

Function/functional form – a function is a description of the relationship that governs the behaviour of two or more variables. The precise mathematical description of this relationship is termed its functional form.

Instrumental variables – a technique often used to overcome the problem of an endogenous explanatory variable. An instrumental variable is used as a proxy for the explanatory variable that is correlated with the error term. The proxy variable to be used as an instrument must be correlated with the explanatory variable in question but not with the error term in the model.

Log-linear model – sometimes called a semi-log model. A log-linear model is a mathematical function, which traces the proportional change in the dependent variable for a given absolute change in the value of the independent variable.

Linear model – a mathematical function, which traces a straight line on a graph.

Measurement error/errors in measurement – the problem of dependent or independent variables that are measured with error. If the problem is errors of measurement in the dependent variable, an OLS regression will still give an unbiased estimate of the parameters, although the estimated variances will be larger than without measurement error. This will generate large confidence intervals. If the problem is errors of measurement in the independent variable, the OLS estimators will be biased and inconsistent (*i.e.* biased even in large samples).

⁸³ Sources include Bannock *et al* (1992) and Gujarati (1995).

Multi-level model – a model that explicitly takes into account the hierarchical structure of data. For example, a multi-level model can take into account that children are clustered in classes, classes are clustered within schools and schools are clustered within LEAs. The model allows random variation between the different levels in the data by including random variables at each level. The fixed or non-random part of the model can also contain explanatory variables measured at each level.

Non-parametric – a technique that does not assume a particular functional form for the underlying model.

Omitted variable bias – the bias to a parameter estimate caused by omitting a variable from the model that should have been included. In multivariate analysis the direction of this bias may not be determined.

Optimum/Optimisation – an optimum is a position in which the primary objective of any economic unit (*e.g.* to maximise revenue) is being served as effectively as it possibly can, within the constraints applying. Individuals and organisations are generally assumed to be rational in economic theory and therefore exhibit optimising behaviour.

Ordinary least squares – a statistical technique for estimating the relationship between a dependent variable and independent variables. Imagine a two-dimensional example, *i.e.* with just one independent variable. The relationship between these two variables can be plotted on an X-Y scatter plot. The least squares regression technique finds the relationship between the variables such that the difference between the actual observations and those traced by a best fit line between the two variables is at a minimum.

Parameter – the values in a mathematical function which remain constant against movements in the variables of the function. In the equation $O=aX + u$, a is a parameter which stays constant as X (the independent variable) and O (the dependent variable) change.

Parametric – a technique that assumes a particular functional form for the underlying model being estimated.

Proxy method – an alternative to Instrumental Variables, which attempts to overcome the endogeneity problem. This strategy assumes that the endogeneity problem arises because of omitted variables, *i.e.* because variables measuring individuals' characteristics (background, attitudes *etc.*) are missing from the model. The proxy method therefore suggests saturating the model with as many explanatory variables as appropriate, in order to control as far as possible for unobserved heterogeneity (*i.e.* differences between individuals). If these explanatory variables do not adequately proxy the unobserved characteristics of individuals then estimates using this technique will remain biased.

Random experiment – an experiment whereby individuals are randomly assigned to a treatment and a non-treatment group. The effects of a particular treatment can be better evaluated with a random experiment since individuals do not get the opportunity to choose whether they undergo the treatment. Hence random experiments should reduce the endogeneity problem.

Regression – a mathematical technique for estimating the parameters of an equation which describes the relationship between the independent and dependent variables.

Returns to scale – the proportionate increase in output resulting from proportionate increases in all inputs. If the inputs are doubled and output less than doubles, the situation is one of decreasing returns to scale. If the inputs are doubled and so is the output, the situation is one of constant returns to scale. If the inputs are doubled and output more than doubles, the situation is one of increasing returns to scale.

Simultaneous equation model – a multi-equation model where there is a two-way or simultaneous relationship between the dependent variable and some of the independent variables. These models have one equation per jointly dependent or endogenous variable.

Stochastic error term – random variable taking positive or negative values.

Stochastic frontier – a random function or frontier (see also **educational production function**).

Stochastic noise term – see **stochastic error term**.

Structural model – the underlying hypothesised relationship between all the variables. It is not always possible to estimate a structural model due to data limitations.

Technical efficiency – this term refers to the efficient production of any product or products, *i.e.* a production process is technically efficient when it is impossible to use less of one input (without using more of another input) to produce a given level of output.

Variance – the variance measures the distribution or spread of the values of a variable around its expected (mean) value.

References

- Akerhielm, K. (1995), 'Does Class Size Matter?', The Economics of Education Review, 14, pp.229-241.
- Altonji, J.G. and Dunn, T.A. (1996), 'Using Siblings to Estimate the Effect of School Quality on Earnings', The Review of Economics and Statistics, 78(4), pp.665-671.
- Angrist, J. and Lavy, V. (1999), 'Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement', Quarterly Journal of Economics, 114(2), pp.533-575.
- Audit Commission (1993), *Two Bs or Not? Schools' and Colleges' A-level Performance*, Audit Commission: London.
- Bannock, G., Baxter, R.E. and Davis, E. (1992), *Dictionary of Economics*, Penguin Books: Middlesex, England.
- Barro, R.J. and Lee, J.W. (1996), 'International Measures of Schooling Years and School Quality', American Economic Review, 86(2), pp.218-223.
- Behrman, J.R. and Birdsall, R. (1983), 'Quality of Schooling: Quantity Alone is Misleading', American Economic Review, 73(5), pp.928-946.
- Behrman, J.R., Rosenweig, M.R., and Taubman, P. (1996), 'College Choice and Wages: Estimates Using Data on Female Twins', The Review of Economics and Statistics, 78(4), pp.672-685.
- Betts, J. (1995), 'Does School Quality Matter? Evidence from the National Longitudinal Survey of Youth', The Review of Economics and Statistics, 77(2), pp.231-250.
- Blatchford, P. and Mortimore, P. (1994), 'Issue of Class Size for Young Children in Schools: What Can We Learn from Research?', Oxford Review of Education, 20, pp.411-428.
- Bound, J., Jaeger, D.A., and Baker, R.M. (1995), 'Problems with Instrumental Variables Estimation when the Correlation between the Instruments and the Endogenous Explanatory Variable is Weak', *Journal of the American Statistical Association*, 90(430), pp.443-451.
- Bound, J. and Loeb, S. (1996), 'The Effect of Measured School Inputs on Academic Achievement: Evidence from the 1920s, 1930s and 1940s Birth Cohorts', The Review of Economics and Statistics, 28(4), pp.653-664.
- Bradley, S. and Taylor, J. (1998), 'The Effect of School Size on Exam Performance in Secondary Schools', Oxford Bulletin of Economics and Statistics, 60(3), pp.291-324.
- Bradley, J., Johns, G. and Millington, J. (1999), 'School Choice, Competition and the Efficiency of Secondary Schools in England', Lancaster University Discussion Paper EC/3.

- Burtless, G. (ed.), (1996), *Does Money Matter? The Effect of School Resources on Student Achievement and Adult Success*, Brookings Institute Press: Washington, D.C.
- Carroll, J.B. (1963), 'A Model of School Learning', *Teachers College Record*, 64, pp.723-733.
- Coleman, J.S., Campbell, E.Q., Hobson, C.J., McPartland, J., Mood, A.M., Weinfield, F.D., and York, R.L. (1966), *Equality of Opportunity*, US Government Printing Office: Washington, D.C.
- Cooper, S.T., and Cohn, E. (1997), 'Estimation of a Frontier Production Function for the South Carolina Educational Process', *Economics of Education Review*, 16(3), pp.313-327.
- Creemers, B.P.M. (1994), *The Effective Classroom*, Cassell: London.
- Creemers, B.P.M., and Reezigt, G.J. (1996), 'School Level Conditions Affecting the Effectiveness of Instruction', *School Effectiveness and School Improvement*, 7(3), pp.197-228.
- Crowther, D., Dyson, A., Millward, A. (1998), *Costs and Outcomes for Pupils with Moderate Learning Difficulties in Mainstream Schools*, Research Report RR89, DfEE: London.
- Currie, J. and Thomas, D. (1995), 'Does Head Start Make a Difference?', *American Economic Review*, 85(3), pp.341-364.
- Currie, J. and Thomas, D. (1998), 'School Quality and the Longer-Term Effects of Head Start', NBER Working Paper No. W6362
- Davie, R., Butler, N. and Goldstein, H. (1972), *From Birth to Seven*, Sage: London.
- Dearden, L., Ferri, J. and Meghir, C. (1997), 'The Effect of School Quality on Educational Attainment and Wages', Institute for Fiscal Studies, Working Paper W98/3
- Department of Education and Science (1983), *School Standards and Spending: Statistical Analysis*, DES: London.
- Department of Education and Science (1984), *School Standards and Spending: A Further Appreciation*, DES: London.
- Dewey, J., Husted, T.A., and Kenny, L.W. (2000), 'The Ineffectiveness of School Inputs: A Product of Misspecification?', *Economics of Education Review*, 19(1), pp.27-45.
- DfEE (1998a), *Developing 16-19 Public Funding Cost Comparisons*, DfEE: London.
- DfEE (1998b), *The Public Funding Costs of Education and Training for 16-19 year olds in England 1996-7*, DfEE: London.
- DfEE (2000), *The Government's Expenditure Plans 2000-01 to 2001-02, Department for Education and Employment and Office for Standards in Education Report*, Stationery Office.

- Dolton, P. and Vignoles, A. (1999), 'The Impact of School Quality on Labour Market Success in the UK', Discussion Paper No.98-03, University of Newcastle.
- Du, J., Green, J. and Peterson, P. (1997), *Effectiveness of School Choice: The Milwaukee Experiment*; Harvard University.
- Dustmann, C., Rajah, N. and van Soest, A. (1998), 'School Quality, Exam Performance and Career Choice', Discussion Paper No. 9816, Tilburg Center for Economic Research.
- Eide, E. and Showalter, M. (1999), 'Factors Affecting the Transmission of Earnings Across Generations: A Quintile Regression Approach', Journal of Human Resources, 34(2), pp.253-267.
- Feinstein, L. and Symons, J. (1999), 'Attainment in Secondary School', Oxford Economic Papers, 51, pp.300-321.
- Fielding, A. (1995), 'Institutional Disparities in the Cost-Effectiveness of GCE A-Level Provision: A Multi-Level Approach', Education Economics, 3(3), pp.249-263.
- Fielding, A. (1998), 'Perspectives on Performance Indicators: GCE Advanced Level and Differences between Institution Types in Cost Effectiveness', *School Effectiveness and School Improvement*, 9(2), pp.218-231.
- Figlio, D.N. (1997a), 'Did the "Tax Revolt" Reduce School Performance?', Journal of Public Economics, 65(3), pp.245-269.
- Figlio, D.N. (1997b), 'Teacher Salaries and Teacher Quality', Economics Letters, 55(2), pp.267-271.
- Figlio, D.N. (1999), 'Functional Form and the Estimated Effects of School Resources', Economics of Education Review, 18(2), pp.241-252.
- Galton, M. and Simon, B. (1980), *Progress and Performance in the Primary Classroom*, Routledge and Kegan Paul: London.
- Ganley, J.A., and Cubbin, J.S. (1992), *Public Sector Efficiency Measurement: Applications of Data Envelopment Analysis*, Elsevier Science: New York.
- Goldhaber, D.D., and Brewer, D.J. (1997), 'Why Don't Schools and Teachers Seem to Matter? Assessing the Impact of Unobservables on Education Production', Journal of Human Resources, 32(3), pp.505-523.
- Goldhaber, D.D., Brewer, D.J. and Anderson, D.J. (1999), 'Three-Way Error Components Analysis of Educational Productivity', Education Economics, 7(3), pp.199-208.
- Goldhaber, D.D., Brewer, D.J., Eide, E.R. and Rees, D.I. (1999), 'Testing for Sample Selection in the Milwaukee School Choice Experiment', Economics of Education Review, 18(2), pp.259-267.
- Goldstein, H. (1987), *Multilevel Models in Educational and Social Research*, Charles Griffin.

- Goldstein, H. (1995), *Multilevel Statistical Modelling*, Edward Arnold: London.
- Goldstein, H. and Blatchford, P. (1998), 'Class Size and Educational Achievement: A Review of Methodology with Particular Reference to Study Design', British Educational Research Journal, 24(3), pp.255-267.
- Greenwald, R., Hedges, L.V. and Laine, R.D. (1996), 'The Effect of School Resources on Student Achievement', Review of Educational Research, 66(3), pp.361-396.
- Gregg, P., and Machin, S. (1999), 'Child Development and Success or Failure in the Youth Labour Market', in D. Blanchflower and R. Freeman (eds.), *Youth Employment and Joblessness in Advanced Countries*, University of Chicago Press.
- Grogger, J. (1996), 'School Expenditures and Post-Schooling Earnings: Evidence from High School and Beyond', *The Review of Economics and Statistics*, 78(4), pp.628-637.
- Gujarati, D.N. (1995), *Basic Econometrics*, McGraw-Hill Inc.: New York.
- Gupta, S., Verhoeven, M. and Tiongson, E. (1999), 'Does Higher Government Spending Buy Better Results in Education and Health Care?' Fiscal Affairs Department: IMF Working Paper WP/99/21.
- Hanushek, E.A. (1986), 'The Economics of Schooling: Production and Efficiency in Public Schools', Journal of Economic Literature, 24(3), pp.1141-1177.
- Hanushek, E.A. (1989), 'The Effect of Differential Expenditures on School Performance', Education Researcher, 18(4), pp.45-51.
- Hanushek, E.A. (1997), 'The Evidence on Class Size', Working Paper No.10, Wallen Wallis Institute of Political Economy, University of Rochester.
- Hanushek, E.A. (1997a), 'Effects of School Resources on Economic Performance', Education Evaluation and Policy Analysis, 19(2), pp.141-164.
- Hanushek, E.A. (1999), 'The Evidence on Class Size', in S.E. Mayer and P.E. Peterson, (eds.), *Earning and Learning: How Schools Matter*, Brooking Institute Press: Washington, D.C., pp.131-168.
- Hanushek, E.A., Rivkin, S.G. and Taylor, L.L. (1996), 'The Identification of School Resource Effects', Education Economics, 4(2), pp.105-125.
- Hanushek, E.A., Kain, J.F., and Rivkin, S.G. (1998), 'Teachers, Schools, and Academic Achievement', NBER Working Paper No. 6691.
- Hanushek, E.A., Kain, J.F. and Rivkin, S.G. (1999), 'Do Higher Salaries Buy Better Teachers?', NBER Working Paper No. 7082.
- Haveman, R.H., and Wolfe, B.L. (1995), 'The Determinants of Children's Attainments: A Review of Methods and Findings', Journal of Economic Literature, 33, pp.1829-1878.

- Heckman, J., Layne-Farrar, A. and Todd, P. (1996a), 'Does Measured School Quality Really Matter? An Examination of the Earnings-Quality Relationship', in G. Burtless (ed.), *Does Money Matter? The Effect of School Resources on Student Achievement and Adult Success*, 1 The Brookings Institute.
- Hedges, L.V., Laine, R.D. and Greenwald, R. (1994), 'Does Money Matter? A Meta-Analysis of Studies of the Effects of Differential Inputs on Student Outcomes', *Educational Researcher*, 23 (April), pp.5-14.
- Hoxby, C.M. (1998), 'The Effects of Class Size and Composition on Student Achievement: New Evidence from Natural Population Variation', NBER Working Paper No. 6869.
- Hoxby, C.M. (1999), 'The Productivity of School and Other Local Public Goods Producers', *Journal of Public Economics*, 74(1), pp.1-30.
- Jesson, D., Mayston, D. and Smith, P. (1987), 'Performance Assessment in the Education Sector: Educational and Economic Perspectives', *Oxford Review of Education*, 13(3), pp.249-266.
- Kirjavainen, T., and Loikkanen, H.A. (1998), 'Efficiency Differences of Finnish Senior Secondary Schools: An Application of DEA and Tobit Analysis', *Economics of Education Review*, 17(4), pp.377-394.
- Krueger, A.B. (1999), 'Experimental Estimates of Education Production Functions', *Quarterly Journal of Economics*, 114(2), pp.497-532.
- Krueger, A.B., and Whitmore, D.M. (1999), 'The Effect of Attending a Small Class in the Early Grades on College-Test Taking and Middle School Test Results: Evidence from Project STAR', Princeton Working Paper No. 427.
- Ladd, H.F. (1999), 'The Dallas School Accountability and Incentive Program: An Evaluation of Its Impacts on Student Outcomes', *Economics of Education Review*, 18(1), pp.1-16.
- Little, A., Mabey, C. and Russell, J. (1973), 'Class Size, Pupil Characteristics and Reading Attainment' in *Literacy at All Levels: Proceedings of the 8th Annual Study Conference*, V. Southgate.
- Lord, R. (1984), *Value for Money in Education*, Public Money: London.
- Marlow, M.L. (2000), 'Spending, School Structure, and Public Education Quality. Evidence from California', *Economics of Education Review*, 19, pp.89-106.
- Marsh, A.J. (2000), 'Resourcing the Continuum of Special Educational Needs in Two LEAs', *Education Management and Administration*, 28(1), pp.77-88.
- Mayston, D.J. (1996), 'Educational Attainment and Resource Use: Mystery or Econometric Misspecification?', *Education Economics*, 4(2), pp.127-142.
- Mayston, D.J. and Jesson, D. (1999), 'Linking Educational Resourcing with Enhanced Educational Outcomes', DfEE Research Report No.179.

- Monk, D.H. (1994), 'Subject Area Preparation of Secondary Mathematics and Science Teachers and Student Achievement', Economics of Education Review, 13, (2), pp.125-145.
- Morris, J. (1959), *Reading in the Primary School*, Newness: London.
- Mortimore, P., Sammons, P., Stoll, L., Lewis, D. and Ecob, R. (1988), *School Matters: the Junior Years*, Open Books: Wells, Somerset.
- Nicolson, R.I., Fawcett, A.J., Moss, H. and Nicolson, M.K. (1999), 'Early Reading Intervention can be Effective and Cost-Effective', Journal of Educational Psychology, 69(1), pp.47-62.
- O'Donoghue, C., Thomas, S., Goldstein, H. and Knight, T. (1997), '1996 Study on Value Added for 16-18 Year Olds in England', DfEE Research Study RS52.
- OFSTED (1995a), *Guidance on the Inspection of Nursery and Primary Schools*, HMSO: London.
- OFSTED (1995b), *Guidance on the Inspection of Secondary Schools*, HMSO: London.
- OFSTED (1996), Effective Sixth Forms, HMSO: London.
- OFSTED (1999a), 'Handbook for Inspecting Secondary Schools', HMSO: London.
- OFSTED (1999b), 'Handbook for Inspecting Nursery and Primary Schools', HMSO: London.
- OFSTED (2000), *The Annual Report of HMCI of Schools*, Stationery Office: London.
- Peterson, P.E., Myers, D.E., Howell, W.G. and Mayer, D.P. (1999), 'The Effects of School Choice in New York Schools', in S.E. Mayer, and P.E. Peterson, *Earning and Learning: How Schools Matter*, Brooking Institute Press: Washington, D.C., pp.317-339.
- Qualifications and Curriculum Authority (1998), *The Long Term Effects of Two Interventions for Children with Reading Difficulties*, QCA: London.
- Reezigt, G.J., Guldmond, H. and Creemers, B.P.M. (1999), 'Empirical Validity for a Comprehensive Model on Education Effectiveness', *School Effectiveness and School Improvement*, 10(2), pp.193-216.
- Reynolds, D., Dammons, P., Stoll, L., Barber, M., and Hillman, J. (1996), 'School Effectiveness and School Improvement in the United Kingdom', School Effectiveness and School Improvement, 7(2), pp.133-158.
- Rouse, C.E. (1998), 'Private School Vouchers and Student Achievement: An Evaluation of the Milwaukee Parental Choice Program', Quarterly Journal of Economics, 113(2), pp.553-602.

- Rouse, C.E. (1999), 'Schools and Student Achievement: More Evidence from the Milwaukee Parental Choice Program', Princeton Working Paper No. 396.
- Ruggiero, J. (1996), 'Efficiency of Educational Production: An Analysis of New York School Districts', The Review of Economics and Statistics, 78(3), pp.499-509.
- Rutter, M., Maughan, B., Mortimore, P., and Ouston, J. (1979), *Fifteen Thousand Hours: Secondary Schools and their Effects on Children*, Open Books: London.
- Scheerens, J. (1997), 'Conceptual Models and Theory-Embedded Principles on Effective Schools', *School Effectiveness and School Improvement*, 8(3), pp.269-310.
- Summers, A.A. and Wolfe, B.L. (1977), 'Do Schools Make a Difference?', American Economic Review, 67, pp.253-267.
- Teddlie, C. and Reynolds, D. (2000), *The International Handbook of School Effectiveness Research*, Fulmer.
- Thanassoulis, E. (1993), 'A Comparison of Regression Analysis and Data Envelopment Analysis as Alternative Methods for Performance Assessments', Journal of Operational Research Society, 44(11), pp.1129-1144.
- Thomas, H. (1990), *Education Costs and Performance*, Cassell.
- Vignoles, A., Desai, T. and Montado, E. (2000), 'The Data Needs of the DfEE Centres for the Economics of Education and the Wider Benefits of Learning', Discussion Paper No. 1, Centre for the Economics of Education, London School of Economics.
- Walberg, H.J. (1984), 'Improving the Productivity of American Schools', Education Leadership, 41, pp.19-27.
- West, A., West, R., Pennell, H. and Travers, T. (1999), 'Financing School Based Education in England: Expenditure, Poverty and Outcomes', Centre for Educational Research, LSE.
- Wiseman, S. (1967), *Children and their Primary Schools: Volume 2*, HMSO: London.
- Witte, J.F., Thorne, C.A. and Sterr, T. (1995), *Fifth Year Report: Milwaukee Parental Choice Program*, Department of Public Instruction: Madison.
- Wright, S.P., Horn, S.P. and Sanders, W.L. (1997), 'Teacher and Classroom Context Effects on Student Achievement: Implications for Teacher Evaluation', Journal of Personal Evaluation in Education, 11, pp.57-67.

CENTRE FOR THE ECONOMICS OF EDUCATION
Recent Discussion Papers

- | | | |
|---|---------------------------------------|--|
| 1 | A. Vignoles
T. Desai
E. Montado | An Audit of the Data Needs of the DfEE Centres for the
Economics of Education and the Wider Benefits of
Learning |
|---|---------------------------------------|--|

To order a discussion paper, please contact the Publications Unit
Tel 020 7955 7673 Fax 020 7955 7595