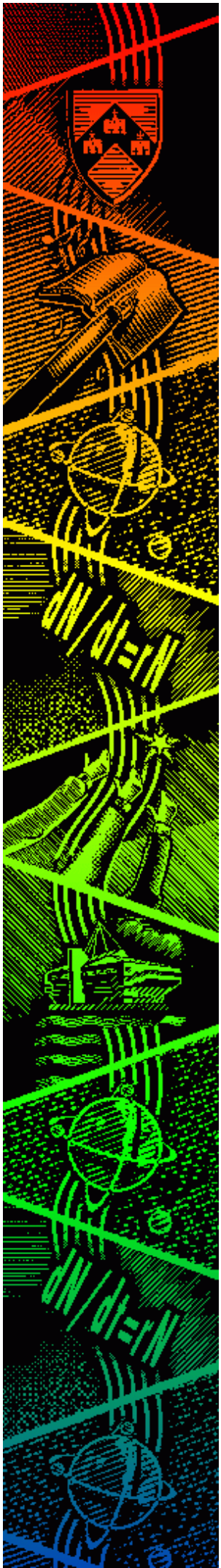

THE UNIVERSITY of York

Discussion Papers in Economics

No. 12/19

The effect of school resources on test scores in England

Cheti Nicoletti and Birgitta Rabe

Department of Economics and Related Studies
University of York
Heslington
York, YO10 5DD

The effect of school resources on test scores in England

Cheti Nicoletti

DERS, University of York and ISER, University of Essex

Birgitta Rabe

ISER, University of Essex

This version: July 14, 2012

Abstract

We analyze the effect of school expenditure on children's test scores at age 16 by means of an education production model. By using unique register data of English pupils, we exploit the availability of test scores across time, subjects and siblings to control for various sources of input omission and measurement error bias. We overcome one of the main criticisms against the value-added model by proposing a novel method to control for the endogeneity of the lagged test. We find evidence of a positive but small effect of per pupil expenditure on test scores.

Keywords: Education production function, cognitive achievements, child development

JEL codes: I22, I24

We thank the Department for Education for making available data from the National Pupil Database. Financial support from the Nuffield Foundation is gratefully acknowledged. The Nuffield Foundation is an endowed charitable trust that aims to improve social well-being in the widest sense. It funds research and innovation in education and social policy and also works to build capacity in education, science and social science research. The Nuffield Foundation has funded this project, but the views expressed are those of the authors and not necessarily those of the Foundation. More information is available at www.nuffieldfoundation.org.

1 Introduction

Increasing amounts of money are being spent on schools around the world, and whether this investment is worthwhile is an important question for policy and parents. However, after long controversy, research is still divided on whether school expenditure has a significant effect on children's cognitive skills. The lack of ideal data and a number of econometric issues are the likely reason for the mixed results found in the empirical literature (e.g. Hanushek et al. 1996; Hanushek 1998; Krueger 2003; Todd and Wolpin 2003).

This paper uses rich administrative data on English state schools to evaluate the effect of school expenditure on pupils' cognitive skills as measured by test scores at age 16, i.e. at the end of compulsory schooling. This evaluation requires the estimation of an education production model with arguments given by past and present inputs by families and schools as well as children's skill endowment. We tackle various issues of input omission as well as measurement error that have plagued the previous literature by estimating different versions of the education production model, each taking account of different sources of estimation bias. Comparison of the different models enables us to assess the relative importance of each bias and to present the first comprehensive evaluation of estimation bias in the schooling quality literature.

The omission of school characteristics and composition may bias estimates of the education production function because of the non-random way in which funding and pupils are allocated across schools. In the English educational system, like in the U.S., the allocation of funding to schools is redistributive, i.e. it is designed to decrease inequalities across children from different backgrounds. If this feature of the allocation of resources is ignored, a positive effect of increasing resources will be understated. The omission of school composition variables also poses an issue because pupils are not randomly distributed across schools and the skill of school mates may affect children's cognitive outcomes. In our education production model we consider a rich set of variables describing school characteristics and composition, including those school-level variables that are used to determine the allocation of funds to schools,[1] the average primary school test scores of peers and the ethnic school-year compo-

[1]This may raise the concern that, after controlling for these funding determinants, we are left with no exogenous variation in school expenditure to be able to identify its effect. But, as we explain in more detail in Section 2.2, exogenous variation is ensured by the fact that the expenditure has increased substantially over the time period considered in our sample, and the rules used to allocate funding to schools have changed

sition, amongst others. Previous U.S. studies have often been unable to control extensively for such school characteristics, and the endogeneity of school resources has been a major methodological difficulty in the schooling quality literature.

The omission of family inputs has also been a major issue in the literature. For example, parents may increase their investment into the child as a result of inadequate spending in school, so that an omission of family characteristics would lead to an overestimation of the school expenditure effect. Unfortunately, family background characteristics are often not available together with school characteristics. In our empirical application we control for family inputs by using sibling fixed effect estimation (Rosenzweig and Wolpin 1994; Altonji and Dunn 1996; Behrmann et al. 1996; Todd and Wolpin 2007). Our data set contains test scores for siblings at the same ages. Sibling estimates take advantage of the fact that some siblings attend different schools and for those attending the same school the inputs of that school will change over time so that the age gap between siblings leads to a different exposure of each sibling to school inputs. Notice that most previous papers on school resources are unable to control appropriately for family characteristics (e.g. Murnane et al. 1981; Hanushek 1986; Hanushek et al. 1996, Dearden et al. 2002; Holmlund et al. 2010).

The omission of past inputs in the education production model is also quite common because of data limitations. A frequently used solution is to adopt a valued added specification, i.e. to include a lagged measure of cognitive skill on the right hand side of the education production model to approximate past inputs (see for example Hanushek 1986; Hanushek et al. 1996). We also adopt a value-added specification in this paper and use test scores at the end of primary school as our measure of past achievements.

The omission of the child skill endowment leads to endogeneity of the lagged cognitive skill in the value added model, however, as unobserved skill endowment is likely correlated with lagged achievement. Therefore the estimation of the valued added model with omitted unobserved child endowment is consistent only if we can accept the assumption that the lagged cognitive skill be independent of the unobserved child skill endowment. This is an assumption that has been criticized but has rarely been relaxed (Todd and Wolpin 2003;

over time, vary regionally and are systematically slow to adapt to changes in a school's need. Identification strategies used in past papers include experiments (e.g. Krueger and Whitemore 2001), instrumental variable approaches and, recently, exploiting boundary discontinuities in school funding (Gibbons et al. 2011).

2007). We propose a new estimation method to take account of this endogeneity. This makes use of the fact that we have lagged and present test scores in three different subjects and consists of two steps. In the first step, we can estimate an individual fixed effects model to control for unobserved individual skill and consistently estimate the effect of lagged achievement. In the second step, we estimate the effect of school resources controlling for the effect of the lagged test and applying school and sibling fixed-effect estimation.

The only other methods used to control for child unobserved endowment using non-experimental data have been dynamic panel data estimation (Todd and Wolpin 2007; Andrabi et al. 2011) and a sort of difference in difference approach which eliminates the child unobserved endowment by considering the difference between adjacent school cohorts in the difference in gains in test scores measured at two different grades (Rivkin et al. 2005). The main advantage of our method over the dynamic panel estimation and the difference in difference approach is that we do not require the education production model, and in particular the coefficient of school inputs, to be invariant across children's ages or grades. This is a quite restrictive assumption (Cunha and Heckman 2007). Therefore dynamic panel estimation works well only if it is based on repeated observations over a very narrow window of the child's life. Furthermore, it is appropriate only if the input of interest has enough individual variation within the narrow age-window considered. Our method provides an alternative solution for cases where these ideal conditions do not hold.

In this paper we also address measurement issues for the cognitive skills and school inputs. We control for measurement error when using exam test scores as proxy for cognitive skill. This is important, as the bias caused by measurement error can exacerbate when using differences in test scores observed for the same pupil in two consecutive years or more in general when adopting panel data estimation (Griliches and Hauman 1986; Bound and Krueger 1991). Similarly, the bias can get magnified when considering the difference between siblings[2] or when the covariates explain a big fraction of the variance of the mis-measured independent variable (Black and Smith 2006). We adopt an instrumental variable approach as in Andrabi et al. (2011), instrumenting lagged test scores in a specific subject using lagged tests in alternative subjects.

[2]This is especially evident when using twins (Ashenfelter and Krueger 1994; Bound and Solon 1999).

We consider how best to measure school inputs. Studies have focused on expenditure per pupil, pupil-teacher ratio, class size, teacher's experience and education - often approximated by teacher's wage (Hanushek 1986). Since we are interested in evaluating the effect of potential changes in secondary school funding on children's cognitive skill, we measure school inputs in terms of expenditure per pupil, which is directly affected by the amount of public funding provided to schools. Expenditure per pupil is mainly determined by spending on teachers and therefore directly related to the pupil-teacher ratio (or class size) and the average teachers' wage, but it can also reflect other types of expenditure such as the cost of teaching resources and of teaching assistants (Holmlund et al. 2010). Studies that focus exclusively on the effect of class size and teacher's wage are unable to capture the effect of a change in school expenditure that operates through channels other than class size and teacher's wage. Furthermore, we assess the potential bias caused by measurement error in the expenditure per pupil by considering averages computed over 1, 3 and 4 years. If there are sporadic school expenditures which fluctuate year on year, we expect that averaging the expenditure per pupil over multiple years reduces the measurement issue.

The main findings of our empirical assessment of the biases caused by omission of inputs and mis-measurement of cognitive skills and expenditure per pupil can be summarized as follows. The omission of school and family characteristics in the education production model causes a large bias of the effect of school expenditure, whereas the omission of unobserved child endowment and the mis-measurement of the cognitive skills do not seem to cause any substantial bias. There seems to be an attenuation bias of the effect of expenditure per pupil when considering expenditure observed in a single year, while 3 or 4-year averages of the expenditure per pupil seem to reduce this bias and provide similar results.

The rest of the paper proceeds as follows. Section 2 gives institutional background on the education and school funding system in England and provides justification of our identification strategy. Section 3 describes the estimation methods and provides formulas for the theoretical asymptotic biases caused by the omission of unobserved child endowment and mis-measurement of the cognitive skills. In Section 4, we describe our data sources and variables used, while in Section 5 we present the estimation results for the education production models and the observed empirical biases. Finally, Section 6 concludes.

2 Institutional background

2.1 Education system in England

Approximately 93% of school children in England attend state schools, the rest are educated in fee-paying private schools. Most schools in England require children to wear a school uniform. Full-time education is compulsory for all children aged between 5 and 16, with most children attending primary school from age 5 to 11 and secondary school from age 11 to 16. The education during these years is divided into four Key Stages, and the National Curriculum sets out targets to be achieved in various subject areas at each of the Key Stages. Pupils undergo externally marked National Curriculum Tests at the end of Key Stages 2 and 4. Until recently such tests were also carried out at Key Stages 1 and 3 but today progress at these stages is examined via individual teacher assessment.

Key Stage 2 National Curriculum Tests are taken at the end of primary school, usually at age 11. Pupils take tests in the three core subjects of English, Mathematics and Science. Key Stage 4 tests are taken at age 16 at the end of compulsory schooling. Pupils enter General Certificate of Secondary Education (GCSE) or equivalent vocational or occupational exams at this stage. They decide which GCSE courses to take, and because English, Mathematics and Science are compulsory study subjects, virtually all students take GCSE examinations in these topics, plus others of their choice, with a total of ten different subjects normally taken. In addition to GCSE examinations, a pupil's final grade may also incorporate coursework elements. Key Stage 2 and 4 test results receive a lot of attention nationally as they play a prominent role in the computation of so-called school league tables, which are used by policy makers to assess schools and by parents to inform school choice.

2.2 School funding in England

This section provides background on how funding was allocated to schools in the time-period 2005-2010 considered in our empirical analysis.[3] The aim is to show that as a result of the

[3]In our empirical analysis we consider test scores of four cohorts of pupils, taking exams in 2007, 2008, 2009 and 2010. School inputs are three-year averages of expenditure per pupil, so that for a student taking exams in 2007, inputs will be from the period 2005-2007. We compare these to averages computed over one and four years

allocation mechanisms used, similar schools can have substantially different funding levels. This is because funding per pupil has increased considerably in real terms from an average of 4,690 pounds in 2005 to 5,750 pounds in 2010 (23% increase in 2010 prices) so that the same school can have differing funding levels over time. Importantly, the rules used to allocate funding across schools have also changed over time, they vary regionally and they are systematically slow to adapt to varying circumstances. These non-linearities are central to our identification strategy.

Most funding for state schools in England comes from central government which hands funds to local education authorities, of which there are 154. The central government grant is calculated mostly on the basis of pupil numbers, deprivation and area costs. The area cost adjustment is intended to adjust for differences in wage costs between areas, although the extra funding received does not generally get passed on to teachers who get paid according to national pay scales. This gives rise to a funding anomaly which Gibbons et al. (2011) exploit to identify the effect of school expenditure on similar schools either side of administrative boundaries. In addition, the so-called spend-plus methodology leads to schools with similar pupil characteristics receiving very different levels of funding. Under this method, local authority grants are determined as flat-rate increases on the grant received the previous year - with a historical starting point in 2005-06 - plus an extra increase based on a formula. "So, current levels of school funding are based on an assessment of needs which is out of date, and on historic decisions about levels of funding which may or may not reflect precisely what schools needed then" (Department for Education 2011, p. 3).

Local authorities then use their own funding formulas to hand out the money received from central government to schools. Apart from pupil numbers, many local authorities historically allow more funding for pupils from deprived backgrounds (eligible for free school meals), with special educational needs and with English as an additional language (Chowdry and Sibieta 2011). There is considerable variation between local authorities in the formula used (West 2009). However, a major constraint that local authorities face when setting their formulas is the Minimum Funding Guarantee introduced in 2004-05 which guarantees each school a minimum increase per pupil per year. Effectively this largely limits the freedom with which local authorities can choose their funding rules. (Levačič 2008). In 2010-11 the

Minimum Funding Guarantee accounted for half the increase in the central school grant (Chowdry and Sibieta 2011).

The combination of spend-plus methodology and Minimum Funding Guarantee has weakened the relationship between school funding levels and educational need. The historical anchor of the funding formula leads to a low reactivity to changes in schools such as rising or falling numbers of deprived pupils. "Some areas are now woefully underfunded compared with how they would be if the system reflected need properly, whereas some areas continue to receive funding to which they should no longer be entitled" (Department for Education 2011, p. 4). In 2010-11 7% of secondary schools had a level of funding at least 10% lower than predicted using observable characteristics, and 6% had funding at least 10% higher (Chowdry and Sibieta 2011, p. 12). These non-linearities are perceived as being too complex and essentially unfair by the current UK government, and reforms to the funding system are being introduced. For the purposes of this paper we can conclude that there is exogenous variation in school expenditure after controlling for pupil and school characteristics.

In Table 1 we give a preview of the between-sibling variation in per pupil expenditure in our data which we describe in more detail in Section 4. This is to show the extent of variation in school expenditure even within the same family, and to demonstrate that this variation is not driven by special groups such as families that tend to move a lot. We find a substantial increase in expenditure per pupil when comparing the two oldest siblings in a family. The mean difference in the expenditure per pupil between siblings attending the same school is 369 pounds (or around 7% of total expenditure) in 2010 prices. For siblings attending different schools - either because of school choice or as a result of a family move[4] - the mean difference is in the region of twice this amount, and these are a minority of siblings.

3 The education production model

The evaluation of the effect of school resources, measured by expenditure per pupil, on children's cognitive test scores would be straightforward if expenditure per pupil was unrelated with other potential determinants of the education production function. But this is not the

[4]Movers are defined here as families that changed address between the exams taken by two siblings in the family.

case because the allocation of resources to schools is redistributive so that school expenditure per pupil is related to school characteristics and pupil composition. Furthermore, family investment in the child cognitive skill may react to changes in the school expenditure, for example compensating for a low school expenditure by increasing private tuition or other cognitive investments.

For this reason we have to consider an education production model that controls not only for school resources but also for all other possible confounding inputs. We focus on cognitive development during the stage that goes from the end of primary school to the end of compulsory schooling, i.e. from about 11 to 16 years of age, and adopt the following education production model:

$$Y^{\star}_{ih,16} = f(I^{F}_{ih}, I^{S}_{ih}, X_{ih}, Y^{\star}_{ih,11}, \mu_{ih}), \tag{1}$$

where $Y^{\star}_{ih,16}$ and $Y^{\star}_{ih,11}$ are unobserved latent cognitive abilities of child i in family h at ages 16 and 11, I^{F}_{ih} is the family investment in the child cognitive development between ages 11 and 16, I^{S}_{ih} is the corresponding school investment, X_{ih} is a vector of other child, household and school characteristics, which are not direct investments in children's cognitive skill but may affect it (e.g. gender, ethnicity, language spoken at home, free school meal eligibility, number of siblings, school characteristics, and pupil composition), and μ_{ih} is the child time-invariant cognitive endowment.

Our estimation sample consists of all pupils enrolled in state schools in England who took their Key Stage 4 tests in the period 2007-2010. For this sample we are unable to observe family and school investments; but we can observe the school expenditure per pupil, which we use as a measure of school investment, and three measures of cognitive abilities each at ages 11 and 16, which are test scores in Mathematics, English and Science obtained in Key Stage 2 and 4 exams. We assume that the relationship between each of these three test scores observed at age 11 and 16 and the unobserved latent cognitive skill at the corresponding age

follows a classical measurement error model[5]

$$Y_{ihs,11} = Y^*_{ih,11} + e_{ihs,11} \text{ and } Y_{ihs,16} = Y^*_{ih,16} + e_{ihs,16}, \tag{2}$$

where the subscript s indicates the test subject and takes value 1 for Mathematics, 2 for English and 3 for Science, $e_{ihs,16}$ and $e_{ihs,11}$ are subject-specific random components identically and independently distributed across children, households and test subjects with mean zero and variance σ^2_e, and are independent of the true latent skill at age 11 and 16, $Y^*_{ih,11}$ and $Y^*_{ih,16}$. The random components $e_{ihs,16}$ and $e_{ihs,11}$ reflect in part a subject specific skill which can persist across time and in part a random error which does not capture any real skill but reflects a measurement error caused for example by inappropriate administration of the subject-specific cognitive test or by temporary variation in the mood and level of attention of a child when taking the test. This implies that while $e_{ihs,16}$ and $e_{ihs,11}$ are identically and independently distributed across children, households and test subjects, they are not independently distributed across time. For this reason, without inconsistency with the classical measurement models (2), we assume that

$$e_{ihs,t} = v_{ihs,t} + \epsilon_{ihs,t}, \tag{3}$$

where t denotes the age of the child and can take value 11 or 16, $v_{ihs,t}$ measures the deviation of the subject-specific latent skill at age t, which we denote $Y^*_{ihs,t}$, from the general latent skill $Y^*_{ih,t}$, and $\epsilon_{ihs,t}$ is a random measurement error.

We assume also that $v_{ihs,t}$ and $\epsilon_{ihs,t}$ satisfy the following conditions:

1. $v_{ihs,t}$ is identically and independently distributed across subjects, children and households with mean zero and variance σ^2_v;

2. $v_{ihs,t}$ is not independently distributed across age and $Cov(v_{ihs,16}, v_{ihs,11}) \neq 0$, whereas there is no correlation across age for different subjects, i.e. $Cov(v_{ihs,16}, v_{ihs',11}) = 0$ if $s \neq s'$;

[5]Imposing a classical measurement error model is equivalent to imposing a factor model with a single factor and equal factor loadings. In Appendix A we report the factor analysis results which seem to confirm that our three school tests are equal to the same latent cognitive skill plus an independent error. The psychologist Spearman (1904) is the pioneer of the factor analysis and he has been the first to apply it to capture a latent measure of skill which he called general intelligence or g-factor. But single factor models, to take account of measurement errors in observed cognitive skill tests, have also been used more recently by economists (e.g. Cunha and Heckman 2008).

3. $\epsilon_{ihs,t}$ is identically and independently distributed across subjects, children, households and age with mean zero and variance σ^2_ϵ;

4. $Cov(\epsilon_{ihs,t}, v_{ihs',t'}) = 0$ for any i, h, s, s', t and t';

5. $v_{ihs,t}$ and $\epsilon_{ihs,t}$ are independent of the true latent skill at age 11 and 16, $Y^*_{ih,11}$ and $Y^*_{ih,16}$, and of the education production function inputs at age 11 and 16 including the unobserved child endowment μ_{ih};

6. the persistence in $Y^*_{ih,t}$, which we define following Andrabi et al. (2011) as the correlation between $Y^*_{ih,16}$ and $Y^*_{ih,11}$ net of the explanatory variables in the education production model, is identical to the persistence in $v_{ihs,t}$.

Under the assumptions defined above and imposing that the production function (1) be additive, separable and linear in its arguments, we can rewrite it as

$$Y_{ihs,16} = \alpha + I^F_{ih}\beta_F + I^S_{ih}\beta_S + X_{ih}\gamma + Y^*_{ih,11}\rho + \mu_{ih} + e_{ihs,16}, \tag{4}$$

where we replaced the unobserved latent cognitive skill at age 16 with the observed test score in subject s and $s = 1, 2, 3$. Model (4) is usually known as the valued added model (see Todd and Wolpin 2003) and it has been extensively used in previous empirical papers to evaluate the contributions of school inputs in a specific stage of the child's school life by controlling for the child's cognitive skill at the beginning of the stage (see Hanushek 1997; Meghir and Rivkin 2011).

3.1 Taking account of measurement error

The mis-measurement of the dependent variable $Y^*_{ih,16}$, under the assumption of a classical measurement error model, does not raise any major concern because it only causes a potential reduction in the precision of our estimation. In contrast, if we replace the unobserved latent cognitive skill at age 11, $Y^*_{ih,11}$, with one of the three observed tests $Y_{ihs,11} = Y^*_{ih,11} + \epsilon_{ihs,11}$, we may bias our estimation results. This is because using $Y_{ihs,11}$ rather than $Y^*_{ih,11}$ the value added model becomes:

$$Y_{ihs,16} = \alpha + I^F_{ih}\beta_F + I^S_{ih}\beta_S + X_{ih}\gamma + Y_{ihs,11}\rho + \mu_{ih} + u_{ihs}, \tag{5}$$

where the new error term $u_{ihs} = e_{ihs,16} - \rho e_{ihs,11}$ is correlated with $Y_{ihs,11} = Y^*_{ih,11} + e_{ihs,11}$. The potential bias caused by measurement error in the estimation of value added models has been emphasized in several papers (see for example Todd and Wolpin 2003), but few of them have considered estimation methods to correct for it. Among the exceptions are Ladd and Walsh (2002) and Andrabi et al. (2011), who consider an instrumental variable approach using instruments given respectively by test scores lagged twice and by alternative test scores, and Cunha and Heckman (2008), who consider a dynamic latent model approach which allows the observed cognitive tests to be related to a latent cognitive skill plus an error.

Notice that Ladd and Walsh (2002), who use the twice lagged test score as instrumental variable, impose the assumption that there is no subject-specific skill transmitted across time, i.e. the assumption that $v_{ihs,t}$ has degenerate distribution. If we want to allow for the more realistic situation where $v_{ihs,t}$ does not have a degenerate distribution, then the estimation using the twice lagged test as instrumental variable is consistent only if the coefficient ρ is equal to the persistence in $v_{ihs,t}$, i.e.

$$\frac{Cov(v_{ihs,16}, v_{ihs,11})}{\sqrt{Var(v_{ihs,16})Var(v_{ihs,11})}} = \frac{Cov(v_{ihs,16}, v_{ihs,11})}{\sigma_v^2} = \rho, \tag{6}$$

or in other words if the persistence of the true latent skill, $Y^*_{ih,t}$, is identical to the persistence of the latent subject-specific skill $Y^*_{ihs,t}$ (see equation (3) above). This assumption seems realistic and is identical to assumption 6 in our above list, and we will make use of it in Sections 3.2 and 3.3. We cannot use the twice lagged test as instrumental variable in this paper, however, as data on this is only available for two of the four academic years included in our estimation sample.

We therefore follow Andrabi et al. (2011) and consider an instrumental variable approach using test scores in alternative subjects as instruments. Using the assumptions of independence of the error term $e_{ihs,11}$ across subjects and the zero correlation between $e_{ihs,16}$ and $e_{ihs',11}$ for $s \neq s'$, we instrument $Y_{ihs,11}$ with $Y_{ihs',11}$ for $s \neq s'$, e.g. we instrument the observed test score in Mathematics at age 11 with the test scores in English and Science at age 11.

The estimation of model (5) using instrumental variables solves the measurement error issue and produces consistent estimates if there are no omitted variables, i.e. if we were able to observe the family and school investments, I^F_{ih} and I^S_{ih}, the child endowment μ_{ih}, the lagged exam tests $Y_{ihs,11}$, and all other individual, family and school characteristics, X_{ih},

that are relevant in the child's cognitive development. Unfortunately, it is generally difficult if not impossible to find a sample with information on all these variables and our sample is no exception.

3.2 Taking account of omitted variables

The omitted variables problem has been one of the major issue in the education production function literature, which has generally been more concerned with the bias caused by omitted school (including class and teacher) characteristics than the one caused by omitted individual and family characteristics. Education production models are usually estimated using datasets with rich information on schools but little or no information on families. Because of these data limitations, most previous papers have failed to control appropriately for child and family characteristics and have controlled for potential omitted school characteristics by considering random or fixed school (or teacher or class) effects estimation (e.g. Goldhaber and Brewer 1997; Steele et al. 2008; Holmlund et al. 2010). Exceptions are papers which evaluate the effect of school inputs using random assignment experiments, discontinuity designs and instrumental variables estimations (e.g. Angrist and Lavy 1999; Krueger 1999; Gibbons et al. 2011), or child fixed effect estimation (e.g. Andrabi et al. 2011; Todd and Wolpin 2007).

While discontinuity design and instrumental variable estimations are applicable only in specific contexts, the child fixed effect estimation requires to observe inputs and test scores in at least three points in the child's life and imposes an age invariant production model, i.e. does not allow the input coefficients to vary across the child's age. This assumption is quite restrictive and previous papers have provided empirical evidence against it. Applied studies have generally found that investments in child's cognitive development are more productive if applied early in life (see Cunha et al. 2006; Cunha and Heckman 2007 and 2008).

In this paper we propose an new estimation method that does not impose age invariance of the coefficients (see below) and we assess whether omitting unobserved child endowment, school and family characteristics may bias the estimation of the effect of school expenditure per pupil.

In our data set we can observe the school investment I_{ih}^S, the lagged exam test $Y_{ihs,11}$, and some individual, family and school characteristics. Let X_{ih}^C, X_{ih}^F and X_{ih}^S be the sub-vectors of X_{ih} containing all child, family and school variables which are relevant for the education production model and let $\gamma = [\gamma^C, \gamma^F, \gamma^S]$ be the corresponding sub-vectors of coefficients. Furthermore, let divide the sub-vectors X_{ih}^C, X_{ih}^F and X_{ih}^S into the observed and unobserved variables, i.e. $X_{ih}^C = [X_{1,ih}^C, X_{2,ih}^C]$, $X_{ih}^F = [X_{1,ih}^F, X_{2,ih}^F]$ and $X_{ih}^S = [X_{1,ih}^S, X_{2,ih}^S]$ where the subscripts 1 and 2 refer to the observed and unobserved variables respectively. Then we can rewrite model (5) as

$$\begin{aligned} Y_{ihs,16} = \alpha + I_{ih}^F\beta_F + I_{ih}^S\beta_S + X_{1,ih}^C\gamma_1^C + X_{2,ih}^C\gamma_2^C + X_{1,ih}^F\gamma_1^F + X_{2,ih}^F\gamma_2^F \\ + X_{1,ih}^S\gamma_1^S + X_{2,ih}^S\gamma_2^S + Y_{ihs,11}\rho + \mu_{ih} + u_{ihs}. \end{aligned} \tag{7}$$

where $I_{ih}^F, X_{2,ih}^C, X_{2,ih}^F, X_{2,ih}^S$ and the child time invariant cognitive endowment μ_{ih} are unobserved.

School fixed effect estimation

The school fixed effect estimation can be easily performed by rewriting model (7) as

$$\begin{aligned} \ddot{Y}_{ihs,16} = \ddot{I}_{ih}^F\beta_F + \ddot{I}_{ih}^S\beta_S + \ddot{X}_{1,ih}^C\gamma_1^C + \ddot{X}_{2,ih}^C\gamma_2^C + \ddot{X}_{1,h}^F\gamma_1^F + \ddot{X}_{2,h}^F\gamma_2^F \\ + \ddot{X}_{1,ih}^S\gamma_1^S + \ddot{X}_{2,ih}^S\gamma_2^S + \ddot{Y}_{ihs,11}\rho + \ddot{\mu}_{ih} + \ddot{u}_{ihs}, \end{aligned} \tag{8}$$

where the double dot denotes the deviation of a variable from the corresponding school mean. This transformation cancels out all time-invariant school characteristics. This is because the average within school is computed considering all students attending the same school and taking their Key Stage 4 exams in any of the years between 2007 and 2010.

Assuming that all unobserved school variables are time invariant, the term $\ddot{X}_{2,ih}^S$ cancels out from the equation (8); but this does not guarantee the consistency of the estimation because there are still unobserved family and child characteristics. A consistent estimation of the effect of school expenditure per pupil would require that either there were no differences in unobserved family and child characteristics (including the child endowment μ_{ih}) between two pupils attending the same school or that these unobserved family and child characteristics were independent of the explanatory variables included in the model. Since these conditions seem quite restrictive, it is unlikely that the school fixed effect estimation is consistent.

Sibling fixed effect estimation

To better take account of unobserved family investments and characteristics we consider family fixed effect estimation. In practice we use the subsample of sibling pairs[6] to estimate model (7) with variables replaced by their differences between siblings, i.e.

$$\begin{aligned}\Delta Y_{hs,16} = {} & \Delta I_h^F \beta_F + \Delta I_h^S \beta_S + \Delta X_{1,h}^C \gamma_1^C + \Delta X_{2,h}^C \gamma_2^C + \Delta X_{1,h}^F \gamma_1^F + \Delta X_{2,h}^F \gamma_2^F \\ & + \Delta X_{1,h}^S \gamma_1^S + \Delta X_{2,h}^S \gamma_2^S + \Delta Y_{hs,11} \rho + \Delta \mu_h + \Delta u_{hs},\end{aligned} \tag{9}$$

where Δ denotes the difference between siblings, e.g. ΔI_h^F denotes the difference in family investment between siblings living in household h. We assume that siblings have equal family investment and that unobserved child, family and school characteristics are identical between siblings, so that ΔI_h^F, $\Delta X_{2,ihs}^C$, $\Delta X_{2,ihs}^F$ and $\Delta X_{2,ih}^S$ cancel out from the model. Consider that we are conditioning on a large set of observed school characteristics and on a number of child and family variables (see section 4), so that the assumption we really impose is that there remain no differences in unobserved child, family and school characteristics between siblings after controlling for the observed explanatory variables. Even if there were remaining differences in unobserved characteristics between siblings, these would not bias the estimation of the effect of our parameter of interest ρ as long as they were uncorrelated with ΔI_h^F.

Under this assumption this leaves us with just one unobserved variable, which is the child endowment μ_{ih}. If we assume that μ_{ih} is given by the sum of a family component that is invariant between siblings and a child specific component, $\mu_{ih} = \mu_h^F + \mu_{ih}^C$, then $\Delta \mu_h^F$ also cancels out and we can rewrite model (9) as

$$\begin{aligned}\Delta Y_{hs,16} = {} & \Delta I_h^S \beta_S + \Delta X_{1,h}^C \gamma_1^C + \Delta X_{1,h}^F \gamma_1^F \\ & + \Delta X_{1,h}^S \gamma_1^S + \Delta Y_{hs,11} \rho + \Delta \mu_h^C + \Delta u_{hs}.\end{aligned} \tag{10}$$

Except for the error term Δu_{hs}, the only other unobserved variable in the right hand side of equation (10) is $\Delta \mu_h^C$. For the consistency of the sibling fixed effect estimation we need this unobserved difference between siblings in the child time invariant endowment $\Delta \mu_h^C$ to be independent of each of the differences in the explanatory variables. While we can reasonably assume that the school inputs do not depend on μ_{ih}^C, it is certainly true that differences in the contemporaneous and lagged tests depend on differences in μ_{ih}^C and this can bias the sibling fixed effect estimation.

Two-step estimation

[6] We consider only the two oldest siblings from each family, see Data section for more details.

To take account of the endogeneity of the lagged test caused by the unobserved child-specific endowment, μ_{ih}^C, we adopt a two-step estimation.

In the first step we use the three contemporaneous tests and the three corresponding lagged tests for each child to estimate a *child fixed effect model*. This allows us to control for the unobserved child specific endowment that is invariant across subjects and to consistently estimate ρ in the value added model (4), at least in absence of measurement error $\epsilon_{ih,t}$ and under the assumption that the persistence of the true latent skill, $Y_{ih,t}^*$, is identical to the persistence of the latent subject-specific skill $Y_{ihs,t}^*$. Nevertheless, this estimation is unable to identify the remaining slope coefficients because the corresponding variables do not vary across the three tests.

In the second step we use the estimated coefficient ρ to compute a new dependent variable $(Y_{ihs,16} - Y_{ihs,11}\rho)$ which we regress on the remaining variables,

$$Y_{ihs,16} - Y_{ihs,11}\rho = \alpha + I_{ih}^F\beta_F + I_{ih}^S\beta_S + X_{ih}\gamma + \mu_{ih} + u_{ihs,16}. \tag{11}$$

For this regression we adopt sibling fixed effect estimation to control for potential unobserved variables that do not vary between siblings.

3.3 Asymptotic bias of the different estimation methods

In the following we report the asymptotic bias for the coefficient of the lagged test, ρ, when it is estimated using the sibling fixed effect estimation without and with instrumental variables and the child fixed effect estimation, i.e. the first step of our two-step estimation. The estimation bias of ρ may transmit to the other coefficients. Generally we expect that an underestimation (overestimation) of the persistence may cause an overestimation (underestimation) of the contribution of the remaining variables, including our variable of interest, the expenditure per pupil.

Sibling fixed effect estimation without instrumental variables

Let us consider the sibling fixed effect estimation without instrumental variables, let W be the vector of explanatory variables in our valued added model (7), which excludes the lagged test and the unobserved child specific endowment, and let $M_{\Delta W}$ be the projection matrix on the space orthogonal to the one generated by the variables ΔW, i.e. the differences in the

variables W between siblings; then it can be proven that the estimation of the lagged test coefficient, $\hat{\rho}_{FFE}$, converges asymptotically to

$$plim\ \hat{\rho}_{FFE} = \rho + \frac{Cov(\Delta\mu_h^C, M_{\Delta W}\Delta Y_{hs,11})}{Var(M_{\Delta W}\Delta Y_{hs,11})} + \frac{Cov(\Delta v_{hs,16}, \Delta v_{hs,11})}{Var(M_{\Delta W}\Delta Y_{hs,11})} - \rho\frac{Var(\Delta e_{hs,11})}{Var(M_{\Delta W}\Delta Y_{hs,11})} \quad (12)$$

The bias caused by the omission of the unobserved child specific endowment μ_{ih}^C is given by the second addend on the right hand side

$$\frac{Cov(\Delta\mu_h^C, M_{\Delta W}\Delta Y_{hs,11})}{Var(M_{\Delta W}\Delta Y_{hs,11})}, \quad (13)$$

which is supposedly positive because the difference in the lagged cognitive test between two siblings, $\Delta Y_{hs,11}$, is generally positively correlated with the difference in their child specific endowment, $\Delta\mu_h^C$, even after controlling for differences in the observed variables W. The asymptotic bias caused by the measurement error is

$$\left[\frac{Cov(\Delta v_{hs,16}, \Delta v_{hs,11})}{Var(M_{\Delta W}\Delta Y_{hs,11})} - \rho\frac{Var(\Delta e_{hs,11})}{Var(M_{\Delta W}\Delta Y_{hs,11})}\right]. \quad (14)$$

This asymptotic bias is zero if both $v_{ihs,t}$ and $\epsilon_{ihs,t}$ have degenerate distribution or, in other words, they are both equal to a constant and have no variance.

Under the condition 6, i.e. the assumption that the persistence in $Y_{ih,t}^*$ is equal to the persistence in $Y_{ihs,t}^*$ (or $v_{ihs,t}$)

$$\frac{Cov(v_{ihs,16}, v_{ihs,11})}{\sqrt{Var(v_{ihs,16})Var(v_{ihs,11})}} = \frac{Cov(v_{ihs,16}, v_{ihs,11})}{\sigma_v^2} = \rho, \quad (15)$$

and using the assumptions imposed for the error terms $v_{hs,11}$ and $v_{hs,16}$ we have also

$$\frac{Cov(\Delta v_{hs,16}, \Delta v_{hs,11})}{Var(\Delta v_{hs,11})} = \rho. \quad (16)$$

This last equality implies that

$$\frac{Cov(\Delta v_{hs,16}, \Delta v_{hs,11})}{Var(\Delta v_{hs,11}) + Var(\Delta\epsilon_{hs,11})} < \rho, \quad (17)$$

so that the asymptotic bias (14) is negative.

In conclusion, the asymptotic biases caused by measurement error and omission of the unobserved child specific endowment have opposite sign and neutralize each other at least in part.

Sibling fixed effect estimation with instrumental variables

Let us consider the family fixed effect estimation with instrumental variables Z used to instrument the lagged test $\Delta Y_{ihs,11}$ and let P_Z be the projection matrix on the space generated by the variables Z, then the estimation of the coefficient of the lagged test, $\hat{\rho}_{FFE,IV}$, converges asymptotically to

$$plim\ \hat{\rho}_{FFE,IV} = \rho + \frac{Cov(\Delta\mu_h^C, M_{\Delta W} P_Z \Delta Y_{hs,11})}{Var(M_{\Delta W} P_Z \Delta Y_{hs,11})}, \tag{18}$$

The bias is caused by the omission of the unobserved child specific endowment μ_{ih}^C and it is positive because the predicted difference in the lagged cognitive test between two siblings, $P_Z \Delta Y_{hs,11}$, is generally positively correlated with the difference in their child specific endowment even after controlling for the sibling difference in the observed variables W.

Child fixed effect estimation

The child fixed effect estimation in our two-step estimation converges to

$$plim\ \hat{\rho}_{CFE} = \frac{Cov(v_{hs,11}, v_{hs,16})}{Var(v_{hs,11} + \epsilon_{hs,11})}. \tag{19}$$

Under the assumption that the persistence of the subject specific skill, $Y^*_{ihs,t}$, be identical to the persistence in the latent skill $Y^*_{ih,t}$ (condition 6),

$$\frac{Cov(v_{hs,11}, v_{hs,16})}{Var(v_{hs,11} + \epsilon_{hs,11})} < \rho. \tag{20}$$

so that the asymptotic bias is negative and cancels out only if there is no measurement error $\epsilon_{hs,11}$.

In conclusion while the family fixed estimation with instrumental variables tends to overestimate ρ, the child fixed estimation under-estimates it. $\hat{\rho}_{CFE}$ and $\hat{\rho}_{FFE,IV}$ provide us with a lower and an upper bound for the coefficient ρ, while $\hat{\rho}_{FFE}$ provides a value which should be between the two. In empirical application we will assess whether the bias in ρ transmits to other parameters.

4 Data

The empirical analysis is based on the National Pupil Database (NPD), which is available from the English Department for Education and has been widely used for education research. The NPD is a longitudinal register dataset for all children in state schools in England, covering roughly 93% of pupils in England. It combines pupil level attainment data with pupil characteristics as they progress through primary and secondary school. Pupil characteristics are collected in annual school censuses and include, for example, age, gender, ethnicity, the pupil's language group and a low-income marker. Pupil-level outcome data during compulsory schooling includes National Curriculum assessments typically taken at ages 7, 11, 14 and 16. These comprise a mixture of teacher-led and test-based assessment depending on the age of the pupils.

The advantage of using the NPD for our analysis is that it allows us (i) to adopt school input measures using information on all state schools and all pupils enrolled in these schools, (ii) to obtain precise estimates of school inputs even when their effect is small, (iii) to identify pupils who are siblings.

Outcome and observed background

Our outcomes of interest are General Certificate of Secondary Education (GCSE) or equivalent vocational test results at the end of compulsory schooling, usually taken at age 16 (Key Stage 4). We focus on GCSEs because they mark the first major branching point in a young person's educational career, and lower levels of GCSE attainment are likely to have a longer term impact on experiences in the adult labour market. We consider Key Stage 4 results in the core subjects English, Mathematics and Science which are directly comparable to test results at the end of primary school. In Key Stage 4 pupils receive a grade for each GCSE course, where pass grades include A*, A, B, C, D, E, F, G. We use a scoring system developed by the Qualifications and Curriculum Authority to transform these grades into a continuous point score[7] which we refer to as the Key Stage 4 score.

We control for lagged cognitive achievement using Key Stage 2 National Curriculum tests taken at the end of primary school, usually at age 11, in the three core subjects of English, Mathematics and Science. In the Key Stage 2 exams, pupils can usually attain a maximum

[7] A pass grade G receives 16 points, and 6 points are added for each unit improvement from grade G.

of 36 points in each subject, but teachers will provide opportunities for very bright pupils to test to higher levels. All test scores are standardised to have a mean of zero and a standard deviation of one.

The NPD annual school census allows identification of a number of individual and family background variables which we use in our empirical models. These include gender of the pupil, a binary variable coding ethnicity (white, black, mixed, Indian, Pakistani/Bangladeshi, Chinese), whether or not the first language spoken at home is English and whether special educational needs have been identified for the child[8]. Moreover, we can identify whether or not a pupil is eligible for free school meals (FSM). FSM eligibility is linked to parents' receipt of means-tested benefits such as income support and income-based job seeker's allowance and has been used in many studies as a low-income marker (see Hobbs and Vignoles 2007 for some shortcomings). We use as family background variable the number of all siblings in the state school system in 2007. This is an approximation to the true number of siblings as it is derived from our matching of pupils at the same address in 2007 and only includes school-age siblings who are in state schools at that point in time. We also include the number of months a pupil is older than an August-born (the youngest in a school cohort) to control for age at test effects. Finally, the NPD contains information on the level of deprivation in the children's residential neighbourhood, assessed by the Income Deprivation Affecting Children Index.

School-level variables

To the NPD we merge school-level expenditure information from Consistent Financial Reporting data sets for 2004-2010. This data set contains details on different types of income and expenditure for each school. Assuming that pupils may benefit from school expenditure not only in their exam year, but also in the preceding years, we consider the average school expenditure over three years rather than yearly expenditure.[9] We test the sensitivity of our results to using alternative measures of expenditure based on a different number of years. Expenditure per pupil is expressed in 2010 prices, calculated using the GDP deflator.

[8]These are pupils with learning difficulties. Those that have been assessed by local education authorities receive a statement which is usually associated additional funding received by the school. There are also pupils identified by the schools as having special needs, but without statement.

[9]Expenditure per pupil excludes capital expenditure such as new construction, but includes expenditure such as learning resources which may benefit pupils for several years.

In addition we add school-level characteristics to the NPD using Schools, Pupils and their Characteristics tables published by the Department for Education (e.g. Department for Education 2010). These tables are derived from the annual school censuses. School-level characteristics include an indicator of whether the school is a community school[10] or not, the number of pupils in the school, single sex schools, and whether the admission to the school is selective. Most selective schools are grammar schools which select pupils by skill at age eleven. We also characterise schools in terms of their pupil composition, using the proportion of pupils that receive free school meals, whose first language is English, that are of white, black, mixed, Indian, Pakistani/Bangladeshi and Chinese ethnicity and that have special educational needs (with and without statements). As for the expenditure we average these variables describing the pupil composition over three years. We also add cohort mean test scores in English, Science and Maths as school-level controls for prior attainment within the school.

Sibling definition

The NPD includes address data, released under special conditions, which allows us to match siblings in the data set. The first year that full address details were collected in the NPD across all pupil cohorts was 2007. Siblings are therefore defined as pupils in state schools aged 4-16 and living together at the same address in January 2007. Siblings that are not school-age, those in independent schools and those living at different addresses in January 2007 are excluded from our sibling definition. Step and half siblings are included if they live at the same address, and we are not able to distinguish them from biological siblings.[11]

Estimation sample

For our analysis we select two samples from the National Pupil Database. The first sample which we call Main Sample includes all pupils (singletons and siblings) that took Key Stage

[10]Community schools are owned, governed and managed by the Local Education Authority, whereas in voluntary aided and voluntary controlled schools as well as in foundation schools some or all of these functions are carried out by other organisations such as the Church of England in faith schools, for example.

[11]The matching of siblings was carried out using 1) postcode and house number/name for addresses with no flat or block number; 2) postcode, house number/name and flat number for addresses without block number; 3) postcode, house number/name, flat and block number; 4) postcode, flat and block number where house number/name was missing. Of the 7.246 million pupil files with address information contained in the 2007 school census, only 4,158 cases had insufficient address information to produce a match using these criteria, and 1,212 cases were dropped where more than ten siblings were identified at an address, and it is possible that they were falsely identified as siblings (false positives).

4 exams in 2007 or in one of the three following years (2008, 2009, 2010). We remove pupils with duplicate data entries or with missing data on any of the background or school-level variables from the dataset. Moreover, we retain only pupils for whom we have non-missing test scores for all outcomes at both Key Stages 2 and 4 which leads to a reduction in sample size of 13%. We also exclude "special schools" that exclusively cater for children with specific needs, for example because of physical disabilities or learning difficulties, as well as schools specifically for children with emotional and/or behavioural difficulties. The remaining sample contains 1,773,323 pupils.

The second sample, which we call Sibling Sample, is obtained by dropping from the Main Sample all singletons, i.e. children who do not have any sibling in the Main Sample, and keeping only the oldest two siblings for each household. This is to avoid having to expand the dataset to include all sibling pair combinations within each household with the risk of over-representing households with a large number of children. The restriction to the two oldest siblings does not lead to any major changes in our results because in the vast majority of cases there are only two siblings living in the same households: only 22,744 pupils (5.2% of siblings) are third or higher order siblings in our observation window 2007-2010. The resulting Sibling Sample includes 429,414 siblings and we use it every time we adopt the family fixed effect estimation in our empirical analysis.

Tables 2 and 3 describe main characteristics of the two samples. The test scores are quite similar in both samples, and they do not vary hugely by subject. Individual characteristics are also broadly identical between the two samples: half of the pupils are male; nine out of ten pupils have English as their first language; 11% are eligible for free school meals and roughly one in six pupils is deemed to have special educational needs. On average there are 1.9 school-age children in every household with at least one pupil taking Key Stage 4 exams over the time-period 2007-2010, and in the Sibling sample this number is naturally higher at 2.6. The Sibling Sample contains more pupils of Pakistani and Bangladeshi ethnicity and fewer whites than the Main Sample.

The bottom panel of Table 3 displays school characteristics. The expenditure per pupil, averaged over three years, is around £5,000 in 2010 prices. Secondary schools are quite large with more than 1,000 pupils in a school on average. The school-level proportions of pupils with free school meal eligibility, ethnicity and English as their first language are comparable

to the individual level means. 4-5% of pupils go to selective schools and the majority of pupils in our sample are educated in community schools.

5 Empirical Results

5.1 Taking account of omitted school and family variables

Our first estimation is a value added model of test scores at the end of compulsory schooling (Key Stage 4), estimated using ordinary least squares (OLS) separately for Maths, English and Science (Table 4, first three panels). As we are interested in the potential effect of a change in school expenditure on child cognitive achievements and we are concerned with measurement error and endogeneity of the lagged test (i.e. the test measured at the end of primary school, Key Stage 2), we report only the coefficients of the expenditure per pupil and the lagged test. Full results for our preferred models are in Appendix B, Tables B1 and B2. In column (1) of Table 4 we control for a set of child and household characteristics but omit school characteristics at this point. Child and household controls include number of school-age siblings in state schools, first language English, ethnicity (white, black, mixed, Indian, Pakistani/Bangladeshi, Chinese), sex, eligible for free school meals, special educational needs (with and without statement), deprivation score of residence, number of months older than August-born in school cohort, and academic year dummies. Table 4 shows a negative and statistically significant effect of per pupil expenditure on test scores in Maths, English and Science at age 16 and a high persistence in cognitive skill.

This result is likely caused by the endogeneity of school inputs because of omitted school characteristics and composition. Therefore in a next step we add school-level characteristics to the model, which are: number of pupils (full time equivalent); proportion of pupils eligible for free school meals, with first language English, with special educational needs (with and without statement), belonging to different ethnic groups (white, black, mixed, Indian, Pakistani/Bangladeshi, Chinese and others), and indicators for community school, selective school and single sex school. As these include the characteristics used to determine allocation of funds to schools from government, we are controlling for the endogeneity of school inputs, identifying the effect of school resources through changes in funding levels over time and non-linearities in the funding formula. The least square estimation now suggests that there

is no effect of the expenditure per pupil on test scores for any of the three tests considered (see column (2) in Table 4, first three panels). The results are still very likely biased because of the restrictive assumptions imposed by the estimation. In the following, we try to relax these assumptions by adopting more suitable estimation methods.

We begin by relaxing the assumption that there are no unobserved school characteristics and we consider a school FE (fixed effect) estimator, which makes use of school identifiers available for each pupil. Because schools with more active leadership, for example, may attract more public funding, failing to control for unobserved school characteristics may lead to an underestimation of the effect of school expenditure. We are therefore not surprised that, after controlling for unobserved time-invariant school characteristics, the effect of expenditure per pupil is positive and statistically significant at standard levels (see column (3) in Table 4, first three panels).[12] We find that an increase in the expenditure per pupil of 1,000 pounds (measured in 2010 prices) leads to an increase in the Mathematics test score of 0.036 standard deviations, in English of 0.028 and in Science of 0.018 standard deviations. In terms of the raw test scores these increases correspond to 0.4 points in Mathematics, 0.3 in English and 0.2 points in Science. As 6 points are needed for an improvement of one grade and there are 8 grades (A*, A, B, C, D, E, F, G), these effects are quite small. Corresponding results found for English primary school pupils in Holmlund et al. (2010), who also control for school fixed effects, are higher with increases of 0.051, 0.040 and 0.050 standard deviations respectively. Possibly the higher effect found in primary schools is explained by the fact that investments in cognitive skill in early childhood are more effective than investments in later childhood and adolescence (see Cunha and Heckman 2007).

The school FE estimation can control for unobserved school characteristics, but it can still be biased because of unobserved family characteristics (e.g. parental investments) which may affect child cognitive development and be correlated with the school variables. Since we are able to identify children living in the same household in 2007, a remedy to take account of unobservable family characteristics is adopting a sibling FE estimation, which boils down to taking differences between siblings in all variables in the simple regression model for test scores. Under the assumption that sibling share the same family characteristics, differencing the dependent and the explanatory variables cancels out their effect.

[12] In the school FE model we omit the time-invariant indicators for community, selective and single sex school.

The new coefficients estimated for the expenditure per pupil are lower but still positive and statistically significant (see column (4) in Table 4).[13] We find a coefficient of 0.022 for Mathematics, 0.018 for English and of 0.017 for Science. This seems to suggest that controlling for school effects, but neglecting to control for family effects may lead to an overestimation of the effect of expenditure per pupil. However, thesibling FE estimation omits potential unobserved time-invariant school characteristics which differ between siblings. Nevertheless, given that the majority of siblings go to the same school (85%, see Table 1), the time-invariant school effect is likely to be very similar for two children living in the same household. Therefore sibling differences are probably widely unaffected by time-invariant school characteristics. Furthermore, our sibling FE model still controls for a large set of time-varying school characteristics, which may differ between two siblings even if they attend the same secondary school.

In the bottom panel of Table 4 (see panel pooled tests) we also report the estimation results for an education model which imposes equality of coefficients across the three tests. Conclusions are very similar to the ones drawn for the models estimated separately for Maths, English and Science. Looking at the results for our preferred estimation, the sibling FE estimation, we find the coefficient of the expenditure per pupil is very close to the ones estimated using separate models for the three tests, with a difference of at most 16%. In contrast the persistence in cognitive skill seems to change more across subjects.

5.2 Taking account of measurement error on lagged test scores

We next correct for the potential bias caused by measurement error on the lagged test by applying the estimation methods considered in Table 4 but instrumenting the lagged test in a specific subject with the lagged tests in the two alternative subjects. The results are shown in Table 5 and we focus the discussion on the sibling FE estimation with instrumental variables (FEIV) which is our preferred estimation in this table.[14] Using instrumental variables, the

[13]Sibling FE model do not use individual-level variables with no or very little variation between siblings (e.g. dummy variables for ethnic groups and first language English) because their effect would not be identified when considering differences between siblings. The child and household variables which we keep in the sibling FE model are: sex, special educational needs (with and without statement), deprivation score of residence, months older than August-born, academic year.

[14]All fixed effect estimations with instrumental variables are performed using Stata command xtivreg2 (Schaffer 2010), while instrumental variable estimation without fixed effects is performed using Stata command ivreg2 (Baum et al. 2007).

coefficient of the lagged test increases substantially and becomes much more similar across the three tests. The results suggest that there is an attenuation bias caused by measurement error and a potential bigger variance in the measurement error for the English and Science tests than for the Maths test. The consequences of this bias seem to have small repercussions on the coefficient of the expenditure per pupil, which decreases only slightly when considering the FEIV estimation compared to the FE estimation.

There is still an assumption imposed by the sibling FEIV which is not credible. This is the assumption that the lagged test be exogenous. Since both the lagged and the contemporaneous test scores are likely to depend on unobserved individual specific endowments, we have an endogeneity issue. The instruments are likely correlated with unobserved pupil endowment and therefore invalid. This is confirmed by the Hansen's J tests (see Table 5) and, as shown in section 3.3, implies an overestimation of the effect of ρ. An overestimation of the lagged cognitive skill may cause an underestimation of the effect of the remaining explanatory variables. For this reason we expect the effect of school expenditure per pupil to be underestimated when using instruments and we interpret our IV estimates of the expenditure per pupil as a lower bound. By comparing Tables 4 and 5 and looking at the endogeneity tests in Table 5 we can conclude that the IV estimation produces significantly different results especially for the effect of the lagged cognitive skill. Encouragingly the effect of the expenditure per pupil does not change much and, as expected, reduces slightly.

5.3 Taking account of unobserved individual endowment

By assuming that the persistence in the subject specific latent skill be equal to the persistence in the latent skill, we can use the repeated observations available for each pupil on the three different contemporaneous and lagged tests to estimate an individual FE model. Under the assumption of no measurement error on the lagged test, i.e. if $\epsilon_{ihs,11} = 0$ for all children, this model allows us to control for the unobserved individual endowment and to estimate consistently the lagged test coefficient. The individual FE estimation does not allow us to identify any of the other effects because school and pupil characteristics do not change across types of test. Nevertheless we can use a two-step procedure to consistently estimate the coefficients for the remaining variables. We consider as new dependent variable the current test minus the lagged test multiplied by its estimated coefficient in the first step and

regress it on the remaining covariates in the second step. As before we consider four types of specification of the education model: without and with observed school controls, with school and with sibling fixed effects. Our preferred estimation is the sibling FE estimation because it allows the explanatory variables to depend on unobserved family endowment. The individual FE estimate of the test persistence (first step) is reported in column (1) of Table 6, while the estimates of the coefficient of the expenditure per pupil using the four alternative estimations (second step) are reported in column (2) to (5) respectively.[15]

The first step estimation produces a much reduced coefficient for the lagged test. While the estimated coefficients of the lagged test in the sibling FE (sibling FEIV) estimation are 0.622 (0.744), 0.480 (0.711) and 0.468 (0.729) respectively for Mathematics, English and Science (see Tables 4 and 5), the estimated lagged test coefficient reduces to 0.303 when controlling for the individual FE, i.e. for unobserved individual skill. This result together with the fact that the individual unobserved component (individual cognitive endowment) explains 66.2% of the variance in the contemporaneous test that is unexplained by the lagged test, tells us that the highly statistical and substantive significance of the lagged test is in part explained by the fact that children with high cognitive endowments are likely to succeed in both Key Stage 2 and 4 test results. In contrast the estimated coefficients for the expenditure per pupil increase only slightly in the two-step estimations. In particular, considering our preferred estimate in Table 6, which is the sibling fixed effect estimate in column (5), the effect of expenditure per pupil is 0.022 which is only about 16% higher than the sibling fixed effect estimate in last column of the bottom panel of Table 4.

Notice that the two-step estimation is consistent only if there is no measurement error on the lagged test. More in general, if we relax this assumption, the two-step estimation produces an underestimated persistence which might in turn cause an overestimation of the expenditure per pupil effect (see Section 3). In conclusion our estimates allow us to say that the persistence in the test is between 0.303 (see individual FE estimation in Table 6) and 0.730 (see sibling FEIV for the pooled tests in Table 5), while the effect of the expenditure per pupil is always above 0.016 and below 0.022.

[15]We do not bootstrap the standard errors to take account of the fact that we replace ρ with its estimate because our first step has very low standard errors and makes use of the universe of pupils, so we do not expect our standard errors to change much.

5.4 Assessing the biases caused by different variable omissions and measurement errors

As emphasized in the methodological section the two-step and IV estimations provide an underestimation and an overestimation of the lagged skill coefficient, repectively. Since we expect that an underestimation (overestimation) of the lagged skill coefficient may cause an overestimation (underestimation) of the remaining inputs, we interpret the two-step and IV estimations of the expenditure per pupil effect as an upper and a lower bound on its true effect. In other words, we are unable to produce a point estimate of the effect of expenditure per pupil, but we can partially identify it and provide an interval estimate.

In the first row of Table 7 we report this interval estimate obtained using our two preferred estimations, i.e. the sibling fixed effect estimation with all control variables and using instrumental variables (see last column in the bottom panel in Table 5) and the two-step estimation that uses sibling fixed effect estimation with all observed control variables in the second step (see last column in Table 6). In the following rows we report the corresponding estimates when (i) omitting to control for the unobserved family effect but considering the unobserved school effects and all observed control variables, (ii) including all observed control variables, (iii) including all observed control variables except school controls. The four different identified intervals do not overlap at all and this allows us to state that omitting school fixed effects causes an overestimation of the expenditure per pupil effect, whereas omitting both school and family characteristics leads to a substantial underestimation. The omission of school and family characteristics (see last two rows of Table 6) bias our estimates considerably and leads to IV estimates which are higher rather than lower than the two-step estimates. In the last two rows, both IV and two-step estimates are well outside the range of possible true values identified by our preferred estimates (0.016, 0.022), so we can strongly reject the education models that omit school and family characteristics.

Furthermore, since the identified range of possible true values for the effect of expenditure per pupil, (0.016, 0.022), is quite small, we can infer that the biases caused by the omission of unobserved child endowment and by measurement error on the lagged test scores are probably smaller than the biases caused by the omission of observed and unobserved school variables.

An issue that we have overlooked so far is the potential measurement error on the expenditure per pupil. Theoretically we would like to consider a measure of expenditure per pupil which reflects long term rather than short term school investments. This is because short term expenditure may include sporadic components which are noisy signals that do not really capture school investments in the pupils' cognitive development. We expect that averaging the expenditure per pupil over multiple years reduces the possible measurement error. To assess this claim, we also consider two alternative measures of expenditure per pupil, (1) using the contemporaneous expenditure in the Key Stage 4 exam year only and (2) using a 4 rather than 3-year average.

In the first row of Table 8 we again report the estimation of the lower and upper bound for the effect of the expenditure per pupil for the 3-year average expenditure (i.e. the FEIV and the two-step estimation with sibling fixed effect and all controls as reported in the first row in Table 7) followed by the equivalent estimates when using 4-year and and 1-year average expenditure per pupil. While the identified intervals for the 4 and 3-year average expenditure are overlapping, the interval identified using 1-year average expenditure does not overlap at all with the other two and it is substantially lower.

The effect of per pupil expenditure is positive and statistically significant when using 1-year expenditure, but the estimated effects are considerably lower than those using 3 and 4-year averages (0.009 in the two-step model using sibling FE and 0.005 in the sibling FEIV model using pooled tests). When using a 4-year average, the effects are only slightly higher at 0.026 and 0.019 respectively. This corroborates our suspicion of bigger measurement error on the yearly expenditure per pupil, which cancel out or at least reduce substantially when considering average of expenditure over multiple years.

6 Conclusions

In this paper, we have used English register data from the National Pupil Database 2007-2010 to investigate the effect of per pupil expenditure on test scores in Mathematics, English and Science at the end of secondary school. Our major finding is that a rise in the expenditure per pupil of £1,000 leads to an increase in the test scores of about 2% of a standard deviation. This effect is statistically significant but very small.

Owing to our unusually rich data, we are able to apply successively a number of estimation techniques that deal with several sources of estimation bias encountered in the previous literature. This allows us to assess which sources of estimation bias are more important than others. The results show that controlling for observed and unobserved school characteristics and controlling for unobserved family investments makes a big difference to the results. Because the majority of siblings are enrolled in the same school, time-invariant unobserved school characteristics are likely to be very similar for two children living in the same household, so that sibling differences should be largely unaffected by them. Therefore we consider the family FE estimation our preferred model.

We also control for the correlation between the individual unobserved endowment and the lagged input by using individual FE estimation. While this exercise shows that the coefficient on the lagged input is heavily upward biased when not controlling for its endogeneity, the effect on the variable of interest - per pupil expenditure - is modest. We also apply a sibling FE estimation with instrumental variables to control for the measurement error on the lagged test. We find that there is a substantial attenuation bias of the persistence but there are little or no repercussions on the coefficient of the expenditure per pupil. Interestingly, even if the sibling FE estimation is biased by the measurement error of the lagged test and by the omission of the unobserved individual endowment, the two biases have opposite signs and seem to cancel each other out.

This result is important for future applications that due to data limitations are forced to use value added models to control for past inputs, and it suggests that the combined bias resulting from endogeneity and measurement error of the lagged test is small. On the other hand, the omission of school and family characteristics causes a large bias in the estimation of the effect of school resources on skills.

References

Altonji J.G. and T.A. Dunn (1996),"Using Siblings to Estimate the Effect of School Quality on Wages" *The Review of Economics and Statistics*, 78(4): 665-671.

Andrabi T., J. Das, A.I. Khwaja and T. Zajonc (2011), "Do Value-Added Estimates Add Value? Accounting for Learning Dynamics" *American Economic Journal: Applied Economics*, 3(3): 29-54

Angrist J. and V. Lavy (1999),"Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievemen" *Quarterly Journal of Economics*, 114(2): 533-575.

Ashenfelter O. and A.B. Krueger (1994), "Estimates of the Economic Return to Schooling from a New Sample of Twins" *American Economic Review*, 84(5): 1157-1173.

Baum, C.F., M.E. Schaffer and S. Stillman (2007), "ivreg2: Stata Module for Extended Instrumental Variables/2SLS, GMM and AC/HAC, LIML and K-class Regression" http://ideas.repec.org/c/boc/bocod

Behrman J.R., M.R. Rosenzweig and P. Taubman, (1996), "College Choice and Wages: Estimates Using Data on Female Twins" *The Review of Economics and Statistics*, 78(4): 672-685.

Black, D.A., J.A. Smith, (2006), "Estimating the Returns to College Quality with Multiple Proxies for Quality" *Journal of Labor Economics*, 24(3): 701-728.

Bound, J. and A.B. Krueger (1991), "The Extent of Measurement Error in Longitudinal Earnings Data: Do Two Wrongs Make a Right?" *Journal of Labor Economics*, 9(1) 1-24.

Bound J. and G. Solon (1999), "Double Trouble: On the Value of Twin-based Estimation of the Return to Schooling" *Economics of Education Review*, 18(2): 2, 169-182.

Chowdry, H. and L. Sibieta (2011), "School Funding Reform: An Empirical Analysis of Options for a National Funding Formula" IFS Briefing Note BN123.

Cunha, F. and J.J. Heckman (2007), "The Technology of Skill Formation" *American Economic Review*, 92(2): 31-47.

Cunha, F. and J.J. Heckman (2008), "Formulating, Identifying and Estimating the Technology of Cognitive and Noncognitive Skill Formation" *Journal of Human Resources*, 43(4): 738-782.

Cunha, F., J.J. Heckman, L.J. Lochner and D.V. Masterov (2006) "Interpreting the Evidence on Life Cycle Skill Formation" in E.A. Hanushek and F. Weich (Eds) *Handbook of the Economics of Education*, chap. 12, Amsterdam: North-Hollland.

Dearden L., J. Ferri and C. Meghir (2002), "The Effect of School Quality on Educational Attainment and Wages" *The Review of Economics and Statistics*, 84: 1-20.

Department for Education (2010), Schools, Pupils and their Characteristics 2010, http://www.education.gov.uk/rs accessed 23.12.2011.

Department for Education (2011): A Consultation on School Funding Reform: Rationale and Principles. http://www.education.gov.uk/consultations/downloadableDocs/School%20Funding %20Reform%20consultation%20final.pdf, accessed 18.1.2012

Gibbons, S., S. McNally and M. Viarengo (2011), "Does Additional Spending Help Urban Schools? An Evaluation Using Boundary Discontinuities" SERC Discussion Paper 90, London School of Economics.

Goldhaber D.D. and D.J. Brewer (1997), "Why Don't Schools and Teachers Seem to Matter? Assessing the Impact of Unobservables on Educational Productivity" *Journal of Human Resources* 32(3): 505-523.

Griliches, Z. and J. Hausman (1986), "Errors in Variables in Panel Data" *Journal of Econometrics*, 31(1): 93-118.

Hanushek E.A. (1986), "The Economics of Schooling: Production and Efficiency in Public Schools" *Journal of Economic Literature*, 24, 1141-1177.

Hanushek E.A., S.G. Rivkin and L.L. Taylor (1996), "Aggregation and the Estimated Effects of School Resources" *The Review of Economics and Statistics* 78: 4, 611-627.

Hanushek E.A. (1997), "Assessing the Effects of School Resources on Student Performance: An Update" *Educational Evaluation and Policy Analysis*, 19: 141-164.

Hanushek, E.A. (1998), "Conclusions and Controversies about the Effectiveness of School Resources" *Economic Policy Review* 4(1), 11-27.

Hobbs G., and A. Vignoles (2007), "Is Free School Meal Status a Valid Proxy for Socio-economic Status (in Schools Research)?" CEEDP, 84. Centre for the Economics of Education, London School of Economics and Political Science, London, UK.

Holmlund, H., S. McNally and M. Viarengo (2010), "Does Money Matter for Schools?" *Economics of Education Review*, 29, 1154-1164.

Krueger A.B. (1999), "Experimental Estimates Of Education Production Functions" *The Quarterly Journal of Economics*, 114(2): 497-532.

Krueger, A.B. and D.M. Whitemore (2001), "The Effect of Attending a Small Class in the Early Grades on College-test Taking and Middle School Test Results: Evidence from Project Star" *Economic Journal*, 111(468): 1-28.

Ladd H.F., and R.P. Walsh (2002), "Implementing Value-Added Measures of School Effectiveness: Getting the Incentives Right." *Economics of Education Review*, 21(1): 1-17.

Levačič, R. (2008), "Financing Schools. Evolving Patterns of Autonomy and Control" *Educational Management Administration & Leadership*, 36(2): 221-234.

Meghir C., and S.G. Rivkin (2011), "Econometric Methods for Research in Education" in (Eds) Hanushek E.A., S. Machin and L. Woessmann, *Handbook of the Economics of Education*, Volume 3, 1-87.

Rivkin S.G., E.A. Hanushek and J.F. Kain (2005), "Teachers, Schools, and Academic Achievement" *Econometrica*, 73(2): 417-458.

Rosenzweig, M. and K.I. Wolpin (1994), "Are there Increasing Returns to the Intergenerational Production of Human Capital? Maternal Schooling and Child Intellectual Achievement" *The Journal of Human Resources*, 29(2): 670-693.

Schaffer, M.E. (2010), "xtivreg2: Stata Module to Perform Extended IV/2SLS, GMM and AC/HAC, LIML and K-class Regression for Panel Data Models" http://ideas.repec.org/c/boc/bocode/s456501.html

Spearman, C.E. (1904), "General intelligence, Objectively Determined And Measured", *American Journal of Psychology*, 15: 201-293.

Steele F., A. Vignoles, and A. Jenkins (2008), "Estimating the Effects of School Resources on Pupil Attainment: A Simultaneous Equations Multi-level Modelling Approac" *Journal of the Royal Statistical Society* Series A, 170(3): 801-824.

Todd, P.E. and K.I. Wolpin (2003), "On the Specification and Estimation of the Production Function for Cognitive Achievement" *Economic Journal* 113 (February), F3-F33.

Todd, P.E. and K.I. Wolpin (2007), "The Production of Cognitive Achievement in Children: Home, School and Racial Test Score Gaps" *Journal of Human Capital*, 1(1): 91-136.

West, A. (2009), "Redistribution and Financing Schools in England under Labour. Are Resources Going Where Needs Are Greatest?" *Educational Management Administration & Leadership* 37(2): 158-179.

Tables

Table 1: Between-sibling variation in per pupil expenditure

	No. sibling pairs	Mean difference in expenditure per pupil
Siblings at same school	182,021	£369
Siblings at different schools (non-movers)	27,293	£645
Siblings at different schools (movers)	5,393	£738
Total	214,707	£414

Notes: National Pupil Database, 2007-2010; Consistent Financial Reporting Data 2005-2010; Schools, Pupils and their Characteristics Data 2005-2010. Pupil expenditure in 2010 prices, calculated using GDP deflator.

Table 2: Descriptive statistics: Unstandardized exam tests scores

	Main Sample		Sibling Sample	
	mean	std. dev.	mean	std. dev.
Key Stage 2 English score	27.0	4.1	27.1	4.1
Key Stage 2 Science score	28.9	3.7	29.1	3.7
Key Stage 2 Maths score	27.4	4.7	27.7	4.6
Key Stage 4 English score	40.2	9.2	40.7	9.1
Key Stage 4 Science score	39.9	10.4	40.6	10.2
KeySstage 4 Maths score	39.1	10.7	40.0	10.6
Number of observations	1,773,323		429,414	

Notes: National Pupil Database, 2007-2010.

Table 3: Descriptive statistics: Explanatory variables

	Main Sample		Sibling Sample	
	mean	std. dev.	mean	std. dev.
Individual characteristics				
Male	0.501		0.503	
No school-age siblings in state schools	1.909	0.920	2.592	0.856
First language English	0.919		0.908	
White	0.853		0.849	
Black	0.031		0.024	
Mixed	0.027		0.023	
Indian	0.024		0.024	
Pakistani/Bangladeshi	0.035		0.049	
Chinese	0.003		0.003	
Other ethnicity	0.028		0.027	
Free school meal eligible	0.106		0.110	
Special educational need, with statement	0.016		0.014	
Special educational need, no statement	0.159		0.152	
Deprivation score of residence	0.203	0.171	0.194	0.168
N months older than August-born	5.471	3.481	5.503	3.479
School characteristics (3 year averages)				
Expenditure per pupil (£/1000)	4.995	0.786	4.967	0.776
Number of pupils (full time equivalent)	1,144	350	1,150	351
Prop. free school meal eligible	0.132	0.112	0.126	0.112
Prop. first language English	0.900	0.172	0.897	0.180
Prop. special educational need, with statement	0.021	0.013	0.021	0.013
Prop. special educational need, no statement	0.165	0.088	0.162	0.087
Prop. white	0.837	0.213	0.835	0.218
Prop. black	0.034	0.078	0.031	0.071
Prop. mixed	0.027	0.025	0.027	0.024
Prop. Indian	0.024	0.067	0.025	0.070
Prop. Pakistani/Bangladeshi	0.036	0.105	0.040	0.117
Prop. Chinese	0.004	0.006	0.004	0.006
Prop. other ethnicity	0.035	0.051	0.035	0.051
Community school	0.577		0.572	
Selective school	0.045		0.048	
Single sex school	0.114		0.113	
KS2 English scores, by cohort	26.96	1.46	27.02	1.47
KS2 Maths scores, by cohort	27.42	1.70	27.47	1.72
KS2 Science scores, by cohort	28.92	1.32	28.97	1.33
Number of observations	1,773,323		429,414	

Notes: National Pupil Database, 2007-2010; Consistent Financial Reporting Data 2005-2010; Schools, Pupils and their Characteristics Data 2005-2010. Pupil expenditure in 2010 prices, calculated using GDP deflator.

Table 4: OLS and fixed effect estimates of the education production function

	OLS No school controls (1)	OLS All controls (2)	School FE All controls (3)	Sibling FE All controls (4)
		Maths		
Expenditure per pupil	-0.018**	0.008	0.036**	0.022**
	(0.005)	(0.005)	(0.009)	(0.004)
Lagged test	0.718**	0.691**	0.691**	0.622**
	(0.002)	(0.002)	(0.002)	(0.002)
		English		
Expenditure per pupil	-0.025**	0.004	0.028**	0.018**
	(0.005)	(0.005)	(0.008)	(0.005)
Lagged test	0.634**	0.594**	0.592**	0.480**
	(0.002)	(0.002)	(0.002)	(0.002)
		Science		
Expenditure per pupil	-0.036**	0.004	0.018+	0.017**
	(0.006)	(0.006)	(0.010)	(0.004)
Lagged test	0.607**	0.575**	0.573**	0.468**
	(0.002)	(0.002)	(0.002)	(0.002)
		Pooled tests		
Expenditure per pupil	-0.026**	0.005	0.027**	0.019**
	(0.005)	(0.005)	(0.006)	(0.003)
lagged test	0.655**	0.620**	0.620**	0.526**
	(0.002)	(0.001)	(0.001)	(0.001)

Notes: + p < .10, * p < .05, ** p < .01. Robust standard errors in parenthesis. Tests are standardized. Robust standard errors in parenthesis. Controls variables include all variables listed in Table 3 plus the standardized lagged test and dummies for academic year. FE stands for fixed effects estimation.

Table 5: Estimates of the education production function using instrumental variables

	IV No school controls (1)	IV All controls (2)	School FE and IV All controls (3)	Sibling FE and IV All controls (4)
		Maths		
Expenditure per pupil	-0.015**	0.008	0.035**	0.020**
	(0.004)	(0.005)	(0.009)	(0.004)
Lagged test	0.820**	0.796**	0.795**	0.744**
	(0.002)	(0.002)	(0.002)	(0.003)
Hansen's J	2.394	75.75	110.0	0.881
	(0.122)	(0.000)	(0.000)	(0.000)
Endogeneity test	2,150	2,083	2,048	4,477
	(0.000)	(0.000)	(0.000)	(0.348)
		English		
Expenditure per pupil	-0.021**	0.001	0.027**	0.013**
	(0.004)	(0.005)	(0.008)	(0.005)
Lagged test	0.818**	0.779**	0.777**	0.711**
	(0.003)	(0.002)	(0.002)	(0.003)
Hansen's J	398.7	352.5	342.0	139.6
	(0.000)	(0.000)	(0.000)	(0.000)
Endogeneity test	1,969	1,871	1,885	8,528
	(0.000)	(0.000)	(0.000)	(0.000)
		Science		
Expenditure per pupil	-0.031**	0.003	0.017+	0.014**
	(0.005)	(0.006)	(0.010)	(0.005)
Lagged test	0.806**	0.770**	0.768**	0.729**
	(0.003)	(0.002)	(0.002)	(0.003)
Hansen's J	1.889	25.72	43.22	8.437
	(0.169)	(0.000)	(0.000)	(0.004)
Endogeneity test	2,375	2,312	2,293	15,684
	(0.000)	(0.000)	(0.000)	(0.000)
		Pooled tests		
Expenditure per pupil	-0.022**	0.004	0.026**	0.016**
	(0.004)	(0.005)	(0.006)	(0.003)
lagged test	0.815**	0.784**	0.781**	0.730**
	(0.002)	(0.002)	(0.001)	(0.002)
Hansen's J	162.5	49.7	59.4	123.2
	(0.000)	(0.000)	(0.000)	(0.000)
Endogeneity test	2,289	2,304	6,421	23,740
	(0.000)	(0.000)	(0.000)	(0.000)

Notes: + p < .10, * p < .05, ** p < .01. Robust standard errors in parenthesis. Tests are standardized. Robust standard errors in parenthesis. Controls variables include all variables listed in Table 3 plus the standardized lagged test and dummies for academic year. FE and IV stand for fixed effects and instrumental variables.

Table 6: Two-step estimation of the education production function

	first step	*second step estimation*			
	individual fixed effect	OLS No school controls	OLS All controls	School FE All controls	Sibling FE All controls
	(1)	(2)	(3)	(4)	(5)
Lagged test	0.303**				
	(0.001)				
Expenditure per pupil		-0.036**	0.008	0.030**	0.022**
		(0.006)	(0.005)	(0.006)	(0.003)
Observations	5,319,969	5,319,969	5,319,969	5,319,969	1,288,242

Notes: + p < .10, * p < .05, ** p < .01. Tests are standardized. Robust standard errors in parenthesis. Controls variables include all variables listed in Table 3 plus the standardized lagged test and dummies for academic year.

Table 7: Estimated lower and upper bound on the effect of expenditure per pupil considering different model specifications

Model specification	Lower bound	(S.E.)	Upper bound	(S.E.)
	(1)	(2)	(3)	(4)
Sibling FE and all controls	0.016**	(0.003)	0.022**	(0.003)
School FE and all controls	0.026**	(0.006)	0.030**	(0.006)
All controls	0.004	(0.005)	-0.008	(0.005)
All controls except school variables	-0.022**	(0.004)	-0.036**	(0.006)

Notes: + p < .10, * p < .05, ** p < .01.

Table 8: Estimated lower and upper bound on the effect of expenditure per pupil measured over different time-periods

	Lower bound	(S.E.)	Upper bound	(S.E.)
	(1)	(2)	(3)	(4)
3-year average expenditure per pupil	0.016**	(0.003)	0.022**	(0.003)
4-year average expenditure per pupil	0.019**	(0.003)	0.026**	(0.003)
1-year average expenditure per pupil	0.005**	(0.002)	0.009**	(0.002)

Notes: + p < .10, * p < .05, ** p < .01.

Appendix A: Factor analysis of the test scores

In table A1 we report the correlations between tests score in Maths, Science and English at key stage 2 and 4, i.e. at about 11 and 16 years. The correlations are high and range from 0.610 to 0.818. The correlation between tests taken in the two different key stages is higher when the two tests are on the same subject, and this is in line with our assumption that $Cov(v_{ihs,16}, v_{ihs,11}) \neq 0$. To understand if the assumptions imposed by the classical measurement error models introduced in Section 3 are acceptable, we carry out a factor analysis separately for the three tests score measures at key stage 2 (age 11) and key stage 4 (age 16). The results reported in Table A2 seem to support the assumption that Maths, English and Science measure a common latent cognitive skill, that one latent factor is enough to explain almost all the total variance, and the relationship between each of the three test scores and this latent factor is very similar with basically identical factor loadings. In conclusion, the factor analysis results do not contradict the assumption of a classical measurement error model for tests scores measured both at stage 2 and key stage 4.

Table A1 Correlations between test scores

	Maths,KS4	Science,KS4	English,KS4	Maths,KS2	Science,KS2	English,KS2
Maths,KS4	1					
Science,KS4	0.818	1				
English,KS4	0.739	0.761	1			
Maths,KS2	0.766	0.673	0.611	1		
Science,KS2	0.672	0.673	0.610	0.790	1	
English,KS2	0.635	0.637	0.705	0.727	0.741	1

Table A2 Results of the factor analysis for key stage 2 and 4 test scores separately

			Key stage 2 tests			
Factor	Eigenvalue	Proportion	Cumulative	*Variable*	Factor1	Uniqueness
Factor1	2.50487	0.835	0.835	Maths	0.9187	0.156
Factor2	0.28589	0.0953	0.9303	English	0.8982	0.1932
Factor3	0.20924	0.0697	1	Science	0.9242	0.1459
			Key stage 4 tests			
Factor	Eigenvalue	Proportion	Cumulative	*Variable*	Factor1	Uniqueness
Factor1	2.54543	0.8485	0.8485	Maths	0.9259	0.1428
Factor2	0.27409	0.0914	0.9398	English	0.9028	0.185
Factor3	0.18048	0.0602	1	Science	0.9345	0.1268

Notes: Eigenvalue is the variance of each factor, Proportion is the proportion of total variance explained by each factor and Cumulative is the corresponding cumulative proportion. Factor1 reports the factor loadings for the first factor. Uniqueness is the proportion of the variance of the specific test which is not explained by the first factor.

Appendix B: Full results

Table B1: Estimated coefficients of expenditure per pupil, lagged test, child and household controls

	Sibling FE IV	Two-step estimation
	(1)	(2)
Lagged test	0.730**	0.303**
	(0.002)	(0.001)
Expenditure per pupil	0.016**	0.022**
	(0.003)	(0.003)
Child and household controls		
Male	-0.096**	-0.087**
	(0.001)	(0.002)
Special educational need, with statement	0.055**	-0.474**
	(0.008)	(0.007)
Special educational need, no statement	-0.094**	-0.343**
	(0.003)	(0.002)
N months older than August-born	-0.008**	0.006**
	(0.000)	(0.000)
Deprivation score of residence	0.001	0.027
	(0.018)	(0.018)
Academic year 2007/08	0.093**	0.083**
	(0.002)	(0.002)
Academic year 2008/09	0.070**	0.054**
	(0.002)	(0.002)
Academic year 2009/10	0.076**	0.054**
	(0.003)	(0.003)

Notes: + p < .10, * p < .05, ** p < .01. Tests are standardized. Robust standard errors in parenthesis. Sibling FE IV refers to the sibling fixed effect model estimated using instrumental variables and assuming equality of the coefficients across the three tests.

Table B2: Estimated coefficients of school controls

	Sibling FE IV (1)	Two-step estimation (2)
School-level controls		
Number of pupils (full time equivalent)	-0.004	0.012+
	(0.007)	(0.007)
Prop. free school meal eligible	-0.466**	-0.477**
	(0.037)	(0.037)
Prop. first language English	-0.113**	-0.111**
	(0.029)	(0.029)
Prop. special educational need, with statement	-0.266+	-0.329*
	(0.151)	(0.149)
Prop. special educational need, no statement	0.058*	0.362**
	(0.023)	(0.023)
Community school	-0.014**	-0.020**
	(0.003)	(0.003)
Selective school	0.178**	0.277**
	(0.010)	(0.010)
Single sex school	0.062**	0.069**
	(0.006)	(0.006)
Prop. white	-0.057+	-0.070*
	(0.032)	(0.032)
Prop. black	0.099+	0.204**
	(0.060)	(0.060)
Prop. mixed	0.047	0.050
	(0.119)	(0.119)
Prop. Indian	0.042	-0.022
	(0.056)	(0.057)
Prop. Pakistani/Bangladeshi	-0.008	0.088+
	(0.051)	(0.050)
Prop. Chinese	0.433+	-0.005
	(0.257)	(0.256)
KS2 English scores, by cohort	-0.005	0.092**
	(0.008)	(0.008)
KS2 Maths scores, by cohort	0.074**	0.109**
	(0.010)	(0.010)
KS2 Science scores, by cohort	-0.124**	-0.034**
	(0.010)	(0.010)

Notes: + p < .10, * p < .05, ** p < .01. Tests are standardized. Robust standard errors in parenthesis. Sibling FE IV refers to the sibling fixed effect model estimated using instrumental variables and assuming equality of the coefficients across the three tests.