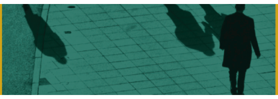
ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistics in Society
Series A



J. R. Statist. Soc. A (2018)
181, *Part* 2, *pp.* 487–515

The effect of school spending on student achievement: addressing biases in value-added models

Cheti Nicoletti

University of York and University of Essex, Colchester, UK

and Birgitta Rabe

University of Essex, Colchester, UK

[Received April 2014. Final revision June 2017]

Summary. The estimation of education production models used to evaluate the effect of school inputs and past skills on test scores, often called value-added models, can be biased by three main econometric issues: unobserved child characteristics, unobserved family and school characteristics and measurement error. We propose a two-step estimation technique which exploits the availability of test scores across time, subjects, families and schools in a unique administrative data set for England to correct for these potential biases. Our empirical results suggest that omitting school characteristics biases the estimation of the effect of school expenditure, whereas omitting unobserved child endowment biases the estimation of the effect of past skills but not the effect of school expenditure.

Keywords: Education production function; School quality; Test scores

1. Introduction

In many countries around the world, schools are spending more money on students than ever before. In the period 2000–2009, expenditure per student increased in each Organisation for Economic Co-operation and Development country by an average of more than 36% (see Organisation for Economic Co-operation and Development (2013).) In England, expenditure per pupil has risen by 69% in real terms over the same period: from £3060 in the year 2000 to £5180 in 2010 (Department for Children, Schools and Families (2009), in 2008 prices). Whether this is a worthwhile use of resources is an important question for policy and parents.

There is a large literature relating public investments in schools to student outcomes in terms of school achievements and qualifications. See Hanushek *et al.* (1996), Krueger (2003), Todd and Wolpin (2003), Hanushek (2006), Meghir and Rivkin (2011) and Gibbons and McNally (2013). The effect of additional expenditure on outcomes of students is often evaluated by using value-added education production functions where child cognitive ability, measured by school test scores, is explained by current inputs and past test scores (e.g. Hanushek (1979, 1986) and Hanushek *et al.* (1996)). Note that the definition of the value-added model that is adopted in this paper should not be confused with the gain score model which explains the gain in test

Address for correspondence: Cheti Nicoletti, Department of Economics and Related Studies, University of York, Heslington, York, YO10 5DD, UK.
E-mail: cheti.nicoletti@york.ac.uk

scores between two school grades or stages by using current inputs and which some references refer to as value added.

Causal estimation approaches usually rely on exogenous variation in school expenditure over time or areas in quasi-experimental designs (e.g. Jenkins *et al.* (2006), Steele *et al.* (2007), Heinesen (2010), Machin *et al.* (2010), Holmlund *et al.* (2010), Gibbons *et al.* (2012), Lavy (2012) and Haegeland *et al.* (2012)). The main econometric issues with estimating such models are input omission and mismeasurement of test scores, which may bias the estimation of the effect of school expenditure on pupil outcomes, of the persistence of achievement between education stages as well as the estimation of other input effects.

This paper assesses the potential biases that are caused by unobserved school, child and family characteristics as well as of measurement error in test scores in estimating the return to school expenditure and the persistence in achievement. We spell out the assumptions that are needed for well-established estimation approaches such as ordinary least squares (OLS) and school fixed effects estimation to yield unbiased estimates. We are especially concerned with the issue of omitted variables when estimating value-added models using school administrative data which typically lack details on family and school characteristics and on children's endowments such as socioemotional abilities and health. The omission of these characteristics can bias both the estimated effect of past test scores and/or of expenditure per student on current test scores. To address this problem we propose a novel two-step estimation procedure. By using administrative data on state schools in England we can compare the coefficients on school expenditure and past achievement estimated by using our two-step procedure with results obtained by using more traditional approaches, thus assessing the magnitude of the resulting biases empirically. The main contributions of this paper therefore are as follows:

(a) to spell out the assumptions that are required for a number of estimation approaches to yield unbiased estimates of both the expenditure per pupil effect and the persistence in achievement;
(b) to propose an estimation technique that accounts for additional sources of unobserved heterogeneity, namely unobserved child characteristics;
(c) to test empirically the importance of controlling for different types of unobserved heterogeneity for the case of England;
(d) to assess the importance of measurement errors in test scores.

Our two-step estimation strategy exploits the availability of test scores in different subjects to control for unobserved child endowments in the first-step estimation and uses school fixed effects to control for unobserved heterogeneity between schools in the second step. The purpose of the first-step estimation is to obtain an unbiased estimate of the persistence of achievement, i.e. the effect of past on current test scores. It is similar to the within-pupil between-subject estimation, which has been used to control for unobserved student characteristics that are invariant across subjects (e.g. Dee (2005, 2007), Clotfelter *et al.* (2010), Slater *et al.* (2010) and Altinok and Kingdon (2012)). By using test scores that are available in different subjects at the end of primary schooling and at the end of compulsory schooling, we can control for unobserved child-specific endowments and evaluate the effect of lagged tests observed at age 11 years on test scores observed at age 16 years ('persistence'). Approaches that have been used to take account of these unobserved endowments by using non-experimental data are dynamic panel data estimation (Todd and Wolpin, 2007; Andrabi *et al.*, 2011) and a difference-in-difference approach which eliminates the unobserved child 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 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 and the effect of omitted childs' endowments, to be invariant across children's ages or grades, which is a restrictive assumption (see Cunha *et al.* (2006), Cunha and Heckman (2007) and Sass *et al.* (2014)).

This child fixed effects estimation similar to ours has been used to estimate the effect of inputs that vary across subjects (such as teacher characteristics and lagged tests) but cannot be used to provide estimates of the effect of explanatory variables that are invariant across subjects such as the school expenditure per pupil. Therefore we introduce a second step. For our second step we use the persistence parameter that is estimated in the first step to generate a new dependent variable: the test score gain between ages 11 and 16 years (this is the age 16 years test score minus the estimated persistence multiplied by the age 11 years test score). We regress this on school expenditure and other control variables and we take account of unobserved school characteristics which can confound the effect of school expenditure by adopting a school fixed effect estimation. Our second-step estimation does not control for unobserved child characteristics, but this is unlikely to bias our results because school expenditure has no variation across pupils and is likely to be independent of pupils' characteristics conditional on our control variables. The second step of our estimation, similarly to Holmlund *et al.* (2010), exploits idiosyncratic variation in expenditure within schools caused by anomalies in funding rules in England for identification (see Section 3.2 for details).

We also address the issue of measurement errors in test scores. Specifically, we adopt an analytic correction method that makes use of reliability ratios of school test scores to derive a correction factor for test scores (Schafer, 1986). We not only take account of errors that are caused by test construction but also of errors that affect school test scores across subjects, e.g. measurement errors caused by the fact that a student was unwell during the examination period. We also implement an alternative correction method based on an approach that was suggested by Boyd *et al.* (2013) and run a sensitivity check to evaluate the consequences for our estimated parameters of considering a much lower reliability ratio than that adopted in the previous two methods.

Our two-step estimation approach can be seen as an extension of the estimation of multilevel value-added models which are usually used to assess the effectiveness of schools, but which we use to estimate the effect of school expenditure (see Aitkin and Longford (1986), Goldstein *et al.* (1993), Ferrão and Goldstein (2009) and Rasbash *et al.* (2010)). Similarly to Rasbash *et al.* (2010) our production model is a multilevel model that allows for the presence of several random effects to take account of unobserved individual (pupil), family, school, neighbourhood and local education authority (LEA) effects. The novelty of our approach is that we take account of

(a) the endogeneity issue caused by the correlation between the lagged test score and the unobserved child effect and between the remaining observed inputs and unobserved school characteristics, and
(b) potential bias caused by measurement errors in test scores.

The results of our two-step estimation show that an increase of £1000 in school spending per student increases test scores by about 6% of a standard deviation. The omission of school characteristics leads to a large underestimation of the effect of expenditure per pupil. In contrast, unobserved child and family background does not lead to a large bias of the expenditure effect once we control for unobserved school characteristics. This is good news for references using administrative data that are unable to control for these characteristics. For researchers who are interested in estimating the persistence of student achievement from one stage of education to the next, however, we find that the omission of child endowment leads to substantial overestimation of the persistence. Similarly, we find that measurement error in test scores does not lead to a large bias of the estimated effect of school expenditure whereas it does lead to an underestimation of persistence.

Even if our proposed two-step estimation of the value-added model takes account of econometric issues that have been neglected by most previous empirical references and contribute to the literature by assessing the potential biases that are caused by such econometric issues, there are some criticizable assumptions which the value-added model imposes (see for a review Boardman and Murnane (1979), Todd and Wolpin (2003), Boyd *et al.* (2013), Lockwood and McCaffrey (2014), Sass *et al.* (2014) and two special issues on ‘Value-added assessment’ published in the *Journal of Educational and Behavioral Statistics* (2004), volume 29, parts 1 and 2. These include linearity and additive separability in inputs, and grade and time invariance of the education production model. We do not assume grade invariance but we discuss the implications of the other assumptions for our empirical application.

The rest of the paper proceeds as follows. Section 2 presents the education production model. Section 2.1 describes the assumptions that are imposed by estimation methods that successively control more extensively for unobserved heterogeneity, starting from OLS estimation which controls only for observed students’ and school characteristics, continuing with school and sibling fixed effect estimations which control additionally for unobserved school and family characteristics, and finally presenting our preferred two-step estimation. Section 2.2 presents our analytic correction method for measurement errors in test scores. Section 3 gives institutional background on the education and school funding system in England and defines the exogenous variation in the expenditure per pupil across time which we exploit to estimate the effect of expenditure. In Section 4, we describe our sources of data and variables that are used whereas in Section 5 we present the estimation results for the education production model, the observed empirical biases and sensitivity checks. Finally, Section 6 concludes.

The programs that were used to analyse the data can be obtained from

```
http://wileyonlinelibrary.com/journal/rss-datasets
```

2. The education production model

We specify our education production model as a value-added model where the child’s cognitive ability is explained by school investments, the child’s past cognitive ability and a set of other control variables. Because the allocation of resources to schools is determined by governmental and local educational authority rules which are redistributive, the amount of expenditure per pupil depends on school characteristics and pupil composition and may differ by LEA. LEAs have responsibility for education within their jurisdiction. For this reason, we must control thoroughly for school and LEA characteristics that might confound the effect of school investment. Furthermore, because both current and past cognitive ability may depend on neighbourhood, family and child characteristics we must control for these additional characteristics to avoid any confounding effect.

We focus on cognitive development during the stage that goes from the end of primary schooling to the end of compulsory schooling in England, i.e. from about 11 to 16 years of age, and adopt the following education production model:

$$Y^*_{ih,16} = f(I^S_{ih}, X_{ih}, Y^*_{ih,11}, \mu_{\text{authority},ih}, \mu_{\text{school},ih}, \mu_{\text{neighbourhood},ih}, \mu_{\text{family},ih}, \mu_{\text{child},ih}, \omega_{ihs,16}), \quad (1)$$

where $Y^*_{ih,16}$ and $Y^*_{ih,11}$ are unobserved latent cognitive abilities of child i in family h at ages 16 and 11 years, I^S_{ih} is the school investment during secondary school up to age 16 years, X_{ih} is a vector of observed child, household and school characteristics, which are not direct investments in children's cognitive skills but proxy for factors that affect them (e.g. gender, ethnicity, language spoken at home, free-school-meal eligibility, number of siblings, school characteristics and pupil composition), $\omega_{ihs,16}$ is a random error which is independent of all other inputs and $\mu_{\text{authority},ih}$, $\mu_{\text{school},ih}$, $\mu_{\text{neighbourhood},ih}$, $\mu_{\text{family},ih}$ and $\mu_{\text{child},ih}$ are unobserved effects which capture all remaining relevant unobserved characteristics at the level of LEA, school, neighbourhood, family and child. We keep the same subscripts for all unobserved effects for simplicity. The assumptions on these unobserved components will depend on the estimation method that is used and we discuss these for several methods below.

To estimate the education production model we have access to administrative data on all pupils enrolled in state schools in England who took their school leaving examinations in the period 2007–2010, and we assume that the model is invariant across the four cohorts of students. We cannot observe family investments in our sample; but we can observe the school expenditure per pupil, which we use as a measure of school investment, and three measures of cognitive abilities at ages 11 and 16 years, which are test scores in mathematics, English and science obtained in National Curriculum examinations. We assume that the relationship between each of these three test scores observed at age 11 and 16 years and the unobserved latent cognitive skill at the corresponding age follows a classical measurement error model

$$\begin{aligned} Y_{ihs,11} &= Y^*_{ih,11} + e_{ihs,11}, \\ Y_{ihs,16} &= Y^*_{ih,16} + e_{ihs,16}, \end{aligned} \tag{2}$$

where the subscript s indicates the test subject and takes value 1 for mathematics, 2 for English and 3 for science, and the subscripts i and h denote children and households respectively. In Section 5 we provide evidence supporting such a type of model. $e_{ihs,11}$ and $e_{ihs,16}$ are subject-specific random components identically and independently distributed across children, households and test subjects with mean 0 and variance σ^2_e, and are independent of the true latent skill at ages 11 and 16 years, $Y^*_{ih,11}$ and $Y^*_{ih,16}$. The random components $e_{ihs,16}$ and $e_{ihs,11}$ in part reflect a subject-specific skill which can persist over time and in part a random error which does not capture any real skill but reflects a measurement error that is caused for example by inappropriate administration of the subject-specific cognitive test or by temporary variation in the level of attention of a child when taking the test. This implies that, although $e_{ihs,16}$ 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 years, $v_{ihs,t}$ measures the deviation at age t of the subject-specific latent skill $Y^*_{ihs,t}$ from the general latent skill $Y^*_{ih,t}$ and $\epsilon_{ihs,t}$ is a random-measurement error.

The assumptions on models (2) and (3) can be restated in terms of $v_{ihs,t}$ and $\epsilon_{ihs,t}$ as the following conditions, which we call maintained assumptions because they are imposed throughout the rest of the paper.

Assumption 1.1. $v_{ihs,t}$ is identically and independently distributed across subjects, children and households with mean 0 and variance σ^2_v.

Assumption 1.2. $v_{ihs,t}$ is not independently distributed across age and $\text{cov}(v_{ihs,16}, v_{ihs,11}) \neq 0$, whereas there is no correlation across age for different subjects, i.e. $\text{cov}(v_{ihs,16}, v_{ihs',11}) = 0$ if $s \neq s'$.

Assumption 1.3. $\epsilon_{ihs,t}$ is identically and independently distributed across subjects, children, households and age with mean 0 and variance σ_ϵ^2.

Assumption 1.4. $\text{cov}(\epsilon_{ihs,t}, v_{ihs',t'}) = 0$ for any i, h, s, s', t and t'.

Assumption 1.5. $v_{ihs,t}$ and $\epsilon_{ihs,t}$ are independent of the true latent skill at age 11 and 16 years, $Y^*_{ih,11}$ and $Y^*_{ih,16}$, and of the education production function inputs at age 11 and 16 years including the unobserved effects.

Assumption 1.6. 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 subject-specific latent skills $Y^*_{ihs,t}$, which implies that the persistence in $v_{ihs,t}$ is identical to the persistence in $Y^*_{ih,t}$.

In Section 5 we assess the validity of these assumptions whenever possible. Under these assumptions and imposing that the production function (1) is additive, separable and linear in its arguments, and replacing the unobserved latent cognitive skill at age 16 and at age 11 years with the observed test score in subject s, we can rewrite model (1) as

$$Y_{ihs,16} = \alpha + I^S_{ih}\beta_S + X_{ih}\gamma + Y_{ihs,11}\rho + \mu_{ih} + u_{ihs}, \tag{4}$$

where $u_{ihs} = e_{ihs,16} - \rho e_{ihs,11} + \omega_{ihs,16}$, and similarly to the multilevel model that was adopted by Rasbash *et al.* (2010):

$$\mu_{ih} = \mu_{\text{authority},ih} + \mu_{\text{school},ih} + \mu_{\text{neighbourhood},ih} + \mu_{\text{family},ih} + \mu_{\text{child},ih}. \tag{5}$$

Of particular interest in this model are the effect of expenditure per pupil and the persistence ρ which measures the self-productivity of the stock of skills at age 11 years. To obtain a consistent estimate of all the parameters of model (4) we need

(a) to control for any unobserved component in μ_{ih} which might be correlated with past test scores or any other control variable and
(b) to correct for the correlation between u_{ihs} and $Y_{ihs,11}$ potentially caused by measurement errors in past test scores.

To obtain a consistent estimate just of our parameters of interest, the effects of expenditure per pupil and the persistence ρ, we do not need to control for the unobserved components of μ_{ih} which are correlated with the control variables as long as the unobserved components are independent of past test scores and expenditure per pupil, conditionally on the control variables.

The parametric assumptions that are imposed by value-added models such as equation (4), in particular the assumptions of invariance of the model across grades and time, and of linearity and additive separability have been criticized. Sass *et al.* (2014) provided empirical evidence that the assumption of invariance across grades is generally rejected. Harris (2007) tested the assumption of linearity in (constant return to) school inputs and found that it cannot be rejected within countries. The assumption of additive separability has been tested among others by Figlio (1999), who showed that productivity of school inputs varies across different levels of student achievements as well as by level of other inputs. Our model does not impose grade invariance but does impose time invariance, linearity and additive separability and we call these assumptions 'parametric functional form assumptions'. We discuss the potential consequences of imposing these in Section 5.4.

The additional assumptions that need to be imposed on the unobserved components μ_{ih} and u_{ihs} in model (4) will depend on the estimation method that is adopted. We shall discuss three methods that have been used in the past to estimate models such as equation (4), OLS estimation, school fixed effect estimation and sibling fixed effect estimation, some of which impose quite restrictive assumptions. We then propose a new two-step estimation method which imposes weaker assumptions. In our empirical analysis we shall show how estimates of the effects of past test scores and school expenditure on current test scores change by imposing increasingly weaker assumptions on unobserved heterogeneity, i.e. on the unobserved components which capture potential omitted variables.

2.1. Taking account of omitted variables

To focus on the issue of omitted variables we assume for the moment that there are no measurement errors in subject-specific test scores, i.e. we assume that the subject-specific latent ability $Y^*_{ihs,t}$ is equal to the observed school test score in subject s, $Y_{ihs,t}$, so that the measurement error $\epsilon_{ihs,t}$ has a degenerate distribution with zero mean and zero variance. This implies that model (4) becomes

$$Y_{ihs,16} = \alpha + I^S_{ih}\beta_S + X_{ih}\gamma + Y_{ihs,11}\rho + \mu_{ih} + \nu_{ihs}, \tag{6}$$

where $\nu_{ihs} = v_{ihs,16} - \rho v_{ihs,11} + \omega_{ihs,16}$ and

$$\mu_{ih} = \mu_{\text{authority},ih} + \mu_{\text{school},ih} + \mu_{\text{neighbourhood},ih} + \mu_{\text{family},ih} + \mu_{\text{child},ih}. \tag{7}$$

2.1.1. *Ordinary least squares estimation with observed school characteristics*

We now turn to specifying assumptions that are required by traditional estimation methods of the education production model, starting with OLS estimation. The consistency of OLS estimation of model (6) requires the following assumptions in addition to the maintained assumptions 1.1–1.6 and the parametric functional form assumptions.

Assumption 2.1. The lagged test, school investment and all other included explanatory variables ($Y_{ihs,11}$, I^S_{ih} and X_{ih}) are uncorrelated with the idiosyncratic error term ν_{ihs}.

Assumption 2.2. $Y_{ihs,11}$, I^S_{ih} and X_{ih} are also uncorrelated with the unobserved child, family, school, LEA and neighbourhood effects or, in short, with the unobserved composite effect μ_{ih}.

If we are interested only in consistently estimating the effects of the school expenditure and the lagged test, rather than all the parameters of the education production model, then a sufficient assumption for the consistency of the estimation of these two effects is the following *conditional independence assumption* (CIA).

Assumption 2.3 (CIA, OLS). Both the idiosyncratic error term ν_{ihs} and the unobserved composite effect μ_{ih} are independent of the lagged test and school investment, $Y_{ihs,11}$ and I^S_{ih}, conditionally on the control variables X_{ih}, i.e. $E[\nu_{ihs} + \mu_{ih}|X_{ih}, Y_{ihs,11}, I^S_{ih}] = E[\nu_{ihs} + \mu_{ih}|X_{ih}]$, and $E[\nu_{ihs} + \mu_{ih}|X_{ih}]$ is linear in the control variables X_{ih}.

The assumption of linearity of the unobserved component, $\nu_{ihs} + \mu_{ih}$, in the control variables can be relaxed if non-parametric rather than OLS estimation was used (see Frölich (2008)).

One of the concerns with the OLS estimations is that unobserved school and LEA characteristics could be correlated with school expenditure per pupil, i.e. we are concerned about the correlation of I_{ih}^S with $\mu_{\text{school},ih}$ and $\mu_{\text{authority},ih}$ in model (6). This correlation can remain even after conditioning on the control variables X_{ih} and therefore it can lead to a biased estimation of the effects of the school investment and lagged test. School fixed effect estimation can correct for this potential issue.

2.1.2. School fixed effect estimation

School fixed effect estimation can be easily performed by transforming the variables in model (6) in deviations from the school mean, i.e. by considering the model

$$\ddot{Y}_{ihs,16} = \ddot{I}_{ih}^S \beta_S + \ddot{X}_{ih}\gamma + \ddot{Y}_{ihs,11}\rho + \ddot{\mu}_{ih} + \ddot{\nu}_{ihs}, \tag{8}$$

where the double dot denotes the deviation of a variable from the corresponding school mean. Because pupils are nested within schools which in turn are nested within LEAs, this transformation cancels out all subject invariant school and LEA characteristics, i.e. the effects $\mu_{\text{school},ih}$ and $\mu_{\text{authority},ih}$, but it does not eliminate the effect of unobserved neighbourhood, family and child characteristics. In our sample we consider pupils from four school cohorts; this implies that we must assume that either unobserved school and LEA characteristics are invariant across the four cohorts or years, or that variation across the four years in unobserved school characteristics are uncorrelated with variation of expenditure per pupil across the four years and lagged test score conditional on the control variables.

The school fixed effect estimation produces consistent estimation of model (6) if the maintained assumptions 1.1–1.6, our parametric functional form assumptions 2.1 and 2.2 and the following additional assumptions hold.

Assumption 3.1. The deviation of the lagged test from its school mean, $\ddot{Y}_{ihs,11}$, is uncorrelated with the corresponding deviation of the idiosyncratic error term, $\ddot{\nu}_{ihs}$.

Assumption 3.2. $\ddot{Y}_{ihs,11}$ is also uncorrelated with the deviation from the school means of the unobserved neighbourhood, family and child effects: $\ddot{\mu}_{\text{neighbourhood},ih}$, $\ddot{\mu}_{\text{family},ihs}$ and $\ddot{\mu}_{\text{child},ih}$.

Assumption 3.3. The deviations of the school investment and all other included explanatory variables from their school mean ($\ddot{I}_{ih}^S$ and $\ddot{X}_{ih}$) are uncorrelated with the corresponding deviation of the idiosyncratic error term $\ddot{\nu}_{ihs}$.

Assumption 3.4. $\ddot{I}_{ih}^S$ and $\ddot{X}_{ih}$ are also uncorrelated with the deviation from the school means of the unobserved neighbourhood, family and child effects: $\ddot{\mu}_{\text{neighbourhood},ih}$, $\ddot{\mu}_{\text{family},ihs}$ and $\ddot{\mu}_{\text{child},ih}$.

If we are interested only in estimating the effects of the school investment and the lagged test, then a sufficient assumption for consistency is the following CIA.

Assumption 3.5 (CIA, school fixed effect). The deviations from the school mean of the idiosyncratic error term $\ddot{\nu}_{ihs}$ and of the unobserved effects $\ddot{\mu}_{\text{neighbourhood},ih}$, $\ddot{\mu}_{\text{family},ihs}$ and $\ddot{\mu}_{\text{child},ih}$ are independent of the corresponding deviations of the lagged test and school investment, $\ddot{Y}_{ihs,11}$ and $\ddot{I}_{ih}^S$, conditional on the control variables $\ddot{X}_{ih}$, i.e.

$$E[\ddot{\nu}_{ihs} + \ddot{\mu}_{\text{neighbourhood},ih} + \ddot{\mu}_{\text{family},ihs} + \ddot{\mu}_{\text{child},ih} | \ddot{X}_{ih}, \ddot{Y}_{ihs,11}, \ddot{I}^S_{ih}]$$
$$= E[\ddot{\nu}_{ihs} + \ddot{\mu}_{\text{neighbourhood},ih} + \ddot{\mu}_{\text{family},ihs} + \ddot{\mu}_{\text{child},ih} | \ddot{X}_{ih}],$$

and $E[\ddot{\nu}_{ihs} + \ddot{\mu}_{\text{neighbourhood},ih} + \ddot{\mu}_{\text{family},ihs} + \ddot{\mu}_{\text{child},ih} | \ddot{X}_{ih}]$ is linear in the control variables $\ddot{X}_{ih}$.

The school fixed effect estimation could be biased because unobserved parental characteristics, $\ddot{\mu}_{\text{family},ih}$, may differ across families within the same school and can be correlated with past test scores of the child, $\ddot{Y}_{ihs,11}$, even after controlling for observable characteristics $\ddot{X}_{ih}$. The endogeneity of $\ddot{Y}_{ihs,11}$ has been emphasized by Todd and Wolpin (2003) who explain that

> 'the value-added formulation [. . .] imposes strong assumptions on the underlying production technology, and the inclusion of a lagged test score as a conditioning variable makes the model highly susceptible to endogeneity bias when data on some of the relevant inputs are missing, even if the omitted inputs are orthogonal to the included inputs'.

The sibling fixed effect estimation corrects for this potential source of endogeneity.

2.1.3. *Sibling fixed effect estimation*

The sibling fixed effect estimation is computed by considering model (6) with variables replaced by their differences between siblings, i.e.

$$\Delta Y_{ihs,16} = \Delta I^S_{ih} \beta_S + \Delta X_{ih} \gamma + \Delta Y_{ihs,11} \rho + \Delta \mu_{ih} + \Delta \nu_{ihs}, \tag{9}$$

where Δ denotes the difference between siblings; for example $\Delta I^S_{ih} = I^S_{ih} - I^S_{i'h}$ denotes the difference in family investment between two siblings (between children i and i' living in the same household h).

Because in our sample siblings belong by definition to the same family, live in the same neighbourhood and go to the same school in the same LEA, the sibling difference transformation cancels out all unobserved effects except for the child effect $\mu_{\text{child},ih}$ and we can rewrite model (9) as

$$\Delta Y_{ihs,16} = \Delta I^S_{ih} \beta_S + \Delta X_{ih} \gamma + \Delta Y_{hs,11} \rho + \Delta \mu_{\text{child},ih} + \Delta \nu_{ihs}. \tag{10}$$

Therefore the consistency of the sibling fixed effect estimation requires the following assumptions.

Assumption 4.1. The difference between siblings in the lagged test, $\Delta Y_{ihs,11}$, is uncorrelated with the corresponding sibling difference in the idiosyncratic error term, $\Delta \nu_{ihs}$.

Assumption 4.2. $\Delta Y_{ihs,11}$ is also uncorrelated with the sibling difference in the unobserved child effect, $\Delta \mu_{\text{child},ih}$.

Assumption 4.3. The differences between siblings in school investment and all other included explanatory variables (ΔI^S_{ih} and ΔX_{ih}) are uncorrelated with the corresponding sibling difference in the idiosyncratic error term, $\Delta \nu_{ihs}$.

Assumption 4.4. ΔI^S_{ih} and ΔX_{ih} are also uncorrelated with the sibling differences in the unobserved child effect, $\Delta \mu_{\text{child},ih}$.

If we are interested only in estimating the effects of the school investment and lagged test, then a sufficient assumption for the consistency is as follows.

Assumption 4.5 (CIA, sibling fixed effect). The differences between siblings in the idiosyncratic error term, $\Delta \nu_{ihs}$, and in the unobserved child effect, $\Delta \mu_{\text{child},ih}$, are independent of the sibling differences in the lagged test score and school investment, $\Delta Y_{ihs,11}$ and ΔI^S_{ih}, conditional on the sibling differences in the control variables ΔX_{ih}, i.e.

$$E[\Delta\nu_{ihs}+\Delta\mu_{\text{child},ih}|\Delta X_{ih},\Delta Y_{ihs,11},\Delta I^S_{ih}]=E[\Delta\nu_{ihs}+\Delta\mu_{\text{child},ih}|\Delta X_{ih}],$$

and $E[\Delta\nu_{ihs}+\Delta\mu_{\text{child},ih}|\Delta X_{ih}]$ is linear in the control variables ΔX_{ih}.

Sibling fixed effect estimation has been used extensively in applied references to control for unobserved family characteristics (Rosenzweig and Wolpin, 1994; Altonji and Dunn, 1996; Behrman *et al.*, 1996; Todd and Wolpin, 2007), and it is consistent when family characteristics are identical between siblings. However, in the context of child cognitive development it is likely that parents invest differentially in two siblings in an attempt either to compensate for or to reinforce differences in their abilities (see Behrman *et al.* (1982), Ermisch and Francesconi (2000) and Bernal (2008)). Therefore, there might be unobserved family characteristics and in particular family investments that differ between siblings. Because the family effect $\mu_{\text{family},ih}$ is by definition identical between siblings, potential parental investment differences between siblings become captured by the sibling difference in the child effect $\Delta\mu_{\text{child},ih}$.

We are concerned about these differences in parental investments because they may be correlated with sibling differences in past test scores, even after controlling for the variables ΔX_{ih}, and this correlation can bias the sibling fixed effect estimation and in particular the estimation of the effects of the lagged test score. Moreover, we are concerned about potential sibling differences in unobserved child-specific characteristics, such as unobserved child socioemotional abilities and health, which can be correlated with sibling differences in past test scores even once we have conditioned on the control variables ΔX_{ih}. To address this we propose a two-step estimation which corrects for the bias that is caused by the potential correlation between $\Delta\mu_{\text{child},ih}$ and $\Delta Y_{hs,11}$.

2.1.4. Two-step estimation

To take account of the endogeneity of the lagged test that is caused by the unobserved child effect, $\mu_{\text{child},ih}$, we adopt a two-step estimation.

In the first step of the two-step estimation procedure we use current test scores in the three subjects English, science and mathematics and the three corresponding past test scores for each child to estimate consistently the persistence parameter by using a *child fixed effect model*. We transform the variables in model (6) in the following way:

$$\tilde{Y}_{ihs,16}=\tilde{Y}_{ihs,11}\rho+\tilde{\nu}_{ihs}, \tag{11}$$

where the tilde over a variable denotes the deviation of the variable from the child mean, i.e. the mean across subjects. Because the expenditure per pupil, I^S_{ih}, the observed explanatory variables X_{ih} and the unobserved effects $\mu_{\text{child},ih}$, $\mu_{\text{family},ih}$, $\mu_{\text{school},ih}$, $\mu_{\text{authority},ih}$ and $\mu_{\text{neighbourhood},ih}$ do not change across subjects, they cancel out from model (11).

The simple regression of $\tilde{Y}_{ihs,16}$ on $\tilde{Y}_{ihs,11}$ provides consistent estimation of ρ under the following assumption.

Assumption 5.1. The deviation of the past test score in subject s from its mean across subjects, $\tilde{Y}_{ihs,11}$, is uncorrelated with the corresponding deviation of the idiosyncratic error term, $\tilde{\nu}_{ihs}$.

In the on-line appendix A we report the asymptotic bias for the coefficient of the lagged test, ρ, when it is estimated by using sibling fixed effect estimation and child fixed effect estimation.

The above first-step estimation is identical to the within-pupil, between-subject estimation that was used by Dee (2005, 2007), the point-in-time fixed effect estimation that was used by

Slater *et al.* (2010) and the student fixed effect estimation that was used by Clotfelter *et al.* (2010). Nevertheless, this estimation cannot identify the remaining slope coefficients—in particular of the expenditure per pupil—because the corresponding variables do not vary across the three tests. Therefore we introduce a second step.

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

$$Y_{ihs,16} - Y_{ihs,11}\hat{\rho} = \alpha + I^S_{ih}\beta_S + X_{ih}\gamma + \mu_{ih} + \nu_{ihs}. \qquad (12)$$

Note that μ_{ih} is not eliminated from the model in the second step, but the parameter ρ is now consistently estimated. For this second-step regression we consider two different types of estimations:

(a) the school fixed effect estimation and
(b) the sibling fixed effect estimation.

The school fixed effect is preferable when the only parameter of interest, besides the persistence that is estimated in the first step, is the effect of the school investment, whereas the sibling fixed effect estimation is preferable when we are interested in the causal effect not only of expenditure per pupil but also of other explanatory variables that might be correlated with unobserved family characteristics (e.g. the effects of free-school-meal eligibility or having special educational needs). In this paper our main parameters of interest are the persistence of the test score and the effect of school investment; therefore our preferred estimation is the two-step estimation with school fixed effect in the second step, but we also consider sibling fixed effect estimation in the second step to show the potential bias in other control variables which could be of interest in other contexts.

2.1.5. *School fixed effect in the second step*

We can implement the school fixed effect estimation in the second step, which controls for potential unobserved school variables, by considering the following transformed model:

$$\ddot{Y}_{ihs,16} - \ddot{Y}_{ihs,11}\hat{\rho} = \ddot{I}^S_{ih}\beta_S + \ddot{X}_{ih}\gamma + \ddot{\mu}_{ih} + \ddot{\nu}_{ihs}, \qquad (13)$$

where as before a double dot denotes the deviation of a variable from its school mean. This school fixed effect estimation enables us to control for unobserved characteristics at the level of LEA and school, $\mu_{\text{authority},ih}$ and $\mu_{\text{school},ih}$, but not for the unobserved neighbourhood, family and child effects $\mu_{\text{neighbourhood},ih}$, $\mu_{\text{family},ih}$ and $\mu_{\text{child},ih}$.

The consistency of the school fixed effect estimation in our two-step procedure requires assumption 5.1 in Section 2.1.4 to hold, as well as assumptions 3.3 and 3.4 in Section 2.1.2, which were also imposed by the school fixed effect estimation. Compared with the school fixed effect estimation our two-step estimation with school fixed effect in the second step allows us to relax the restrictive assumption 3.2 in Section 2.1.2 by allowing for correlation between the unobserved child effect $\ddot{\mu}_{\text{child},ih}$ and the lagged test score expressed as deviations from their school mean.

If we are interested only in estimating the effect of school investment, apart from the persistence in the test score, a sufficient condition for the consistency of the estimation is the following CIA.

Assumption 6.1 (CIA, school fixed effect 2). The deviations from the school mean of the idiosyncratic error term $\ddot{\nu}_{ihs}$ and of the unobserved effects $\ddot{\mu}_{\text{neighbourhood},ih}$, $\ddot{\mu}_{\text{family},ihs}$ and $\ddot{\mu}_{\text{child},ih}$ are independent of the corresponding deviation of the school investment $\ddot{I}^S_{ih}$ conditional on the control variables $\ddot{X}_{ih}$, i.e.

$$E[\ddot{\nu}_{ihs} + \ddot{\mu}_{\text{neighbourhood},ih} + \ddot{\mu}_{\text{family},ihs} + \ddot{\mu}_{\text{child},ih} | \ddot{X}_{ih}, \ddot{I}^S_{ih}]$$
$$= E[\ddot{\nu}_{ihs} + \ddot{\mu}_{\text{neighbourhood},ih} + \ddot{\mu}_{\text{family},ihs} + \ddot{\mu}_{\text{child},ih} | \ddot{X}_{ih}],$$

and $E[\ddot{\nu}_{ihs} + \ddot{\mu}_{\text{neighbourhood},ih} + \ddot{\mu}_{\text{family},ihs} + \ddot{\mu}_{\text{child},ih} | \ddot{X}_{ih}]$ is linear in the control variables $\ddot{X}_{ih}$.

The variation in school investment across time and schools depends on school characteristics, such as the proportion of children who are eligible for free school meals, which are related to the allocation rule of resources across schools (see Section 3.2). Conditionally on school characteristics and school fixed effects, any residual variation in school investment should not depend on unobserved neighbourhood, family and pupil characteristics, which suggests that the assumption of conditional independence, assumption 6.1 (school fixed effect 2), is likely to hold.

2.1.6. *Sibling fixed effect in the second step*

The sibling fixed effect estimation in the second step enables us to control for potential unobserved variables that do not vary between siblings and can be implemented by considering the following transformed model:

$$\Delta Y_{ihs,16} - \Delta Y_{ihs,11} \hat{\rho} = \Delta I^S_{ih} \beta_S + \Delta X_{ih} \gamma + \Delta \mu_{ih} + \Delta \nu_{ihs}, \tag{14}$$

where Δ denotes the difference between siblings (between children i and i' living in the same household h).

Because for the estimation of this sibling fixed effect we use the sample of sibling pairs who live in the same household and neighbourhood and go to the same school in the same LEA, all unobserved effects cancel out from model (14) except for the child effect $\mu_{\text{child},ih}$ so that $\Delta \mu_{ih} = \Delta \mu_{\text{child},ih}$.

The consistency of sibling fixed effect estimation in our two-step procedure requires assumption 5.1 in Section 2.1.4 to hold, as well as assumptions 4.3 and 4.4 in Section 2.1.3, which were also imposed by the sibling fixed effect estimation. Compared with sibling fixed effect estimation our two-step estimation with sibling fixed effect in the second step enables us to relax the restrictive assumption 4.2 in Section 2.1.3 by allowing for correlation between sibling differences in the unobserved child effect $\Delta \mu_{\text{child},ih}$ and the lagged test score.

If we are interested only in estimating the effect of school investment and the persistence in the test score, then sufficient conditions for consistency would be assumption 5.1 and the following CIA.

Assumption 7.1 (CIA, sibling fixed effect 2). The differences between siblings in the idiosyncratic error term $\Delta \nu_{ihs}$ and in the unobserved child effect $\Delta \mu_{\text{child},ih}$ are independent of the sibling differences in the school investment, ΔI^S_{ih}, conditional on the sibling differences in the control variables ΔX_{ih}, i.e.

$$E[\Delta \nu_{ihs} + \Delta \mu_{\text{child},ih} | \Delta X_{ih}, \Delta I^S_{ih}] = E[\Delta \nu_{ihs} + \Delta \mu_{\text{child},ih} | \Delta X_{ih}],$$

and $E[\Delta \nu_{ihs} + \Delta \mu_{\text{child},ih} | \Delta X_{ih}]$ is linear in the control variables ΔX_{ih}.

The standard errors of either of our two-step procedures need to be adjusted to take account of the fact that in the second step ρ is replaced by its estimated value from the first step. To correct for this bias we bootstrap the standard errors by using 50 replications. Note that the two-step estimation is not efficient, but given our sample size of more than 1 million observations we are not concerned about the potential loss of efficiency and we use it as our preferred estimation.

2.2. *Taking account of measurement error*

We now return to the issue of measurement error, examining how we can address measurement error in the observed subject-specific test score when adopting our two-step estimation. Recall that model (4) was

$$Y_{ihs,16}=\alpha+I_{ih}^{S}\beta_S+X_{ih}\gamma+Y_{ihs,11}\rho+\mu_{ih}+u_{ihs}, \tag{15}$$

where $u_{ihs}=e_{ihs,16}-\rho e_{ihs,11}+\omega_{ihs}$, $e_{ihs,t}=v_{ihs,t}+\epsilon_{ihs,t}$ and $Y_{ihs,t}=Y_{ih,t}^{*}+v_{ihs,t}+\epsilon_{ihs,t}$ for $t=11$ and 16 years. Whereas the error $\epsilon_{ihs,16}$ in the left-hand side variable $Y_{ihs,16}$ causes a decrease in the estimation efficiency but no inconsistency, the error $\epsilon_{ihs,11}$ in the lagged test $Y_{ihs,11}$ causes an attenuation bias for the ρ-coefficient (estimated in the first step by using child fixed effect estimation) and a possible overestimation of the effect of the remaining explanatory variables in the second step (school fixed effect estimation). See the on-line appendix A for the bias formula.

To correct for the resulting bias of the child fixed effect estimation of ρ, we multiply the ρ-coefficient that is estimated in the first step by the following correction factor:

$$\text{var}(v_{ihs,11}+\epsilon_{ihs,11})/\text{var}(v_{ihs,11}). \tag{16}$$

This is the so-called analytic correction for measurement error.

We do not observe the correction factor (16), but we can compute it by using information on the reliability ratio $\text{var}(Y_{ih,11}^{*}+v_{ihs,11})/\text{var}(Y_{ih,11}^{*}+v_{ihs,11}+\epsilon_{ihs,11})$, and on the share of the variance of the observed test score in subject s explained by the latent ability $Y_{ih,11}^{*}$, i.e. $\text{var}(Y_{ih,11}^{*})/\text{var}(Y_{ih,11}^{*}+v_{ihs,11}+\epsilon_{ihs,11})$. This is because, under our maintained assumptions 1.1–1.6 there is no correlation between $Y_{ih,11}^{*}$, $v_{ihs,11}$ and $\epsilon_{ihs,11}$, and $\text{var}(Y_{ihs,11})=\text{var}(Y_{ih,11}^{*})+\text{var}(v_{ihs,11})+\text{var}(\epsilon_{ihs,11})=1$. $\text{var}(Y_{ihs,11})=1$ because our test scores are standardized by subject.

He *et al.* (2013) computed the reliability ratios for science, mathematics and English in National Curriculum examinations at the end of primary schooling by using each of the item questions of the primary school tests administered in 2009 in England and found ratios of 0.928, 0.968 and 0.910 for science, mathematics and English respectively. For a recent application of a bias correction based on reliability ratios of cognitive test scores, see Lindqvist and Vestman (2011); for a comparison of the analytic correction method with other methods see Schafer (1986) and Lockwood and McCaffrey (2014). Similar analytic corrections have also been considered by Fuller (1986) and Meyer (1999).

By implementing factor analysis for the three observed lagged test scores, we find that the first factor explains on average 77.5% of the variance of the subject-specific test scores at age 11 years. By considering this common factor as a measure of the latent ability $Y_{ih,11}^{*}$, we can impute to $\text{var}(Y_{ih,11}^{*})/\text{var}(Y_{ih,11}^{*}+v_{ihs,11}+\epsilon_{ihs,11})$ a value of 0.775, which is the average of the share of variance explained by the common factor across the three observed test scores at age 11 years.

By imposing a reliability ratio of 0.935, which is the average across the three very similar ratios that were observed for the three subject-specific test scores in He *et al.* (2013), and $\text{var}(Y_{ih,11}^{*})/\text{var}(Y_{ih,11}^{*}+v_{ihs,11}+\epsilon_{ihs,11})=0.775$, we can assume that the correction factor $\text{var}(v_{ihs,11}+\epsilon_{ihs,11})/\text{var}(v_{ihs,11})$ takes value 1.403.

Our analytic correction method takes account not only of the errors that are caused by test construction but also of errors which similarly affect the test scores in the three subjects, e.g. errors that are caused by the fact that the student may have been unwell or was having trouble at home during the examination period. This is because we consider deviation of each subject-specific test score from the test score averaged across subjects. Therefore any error that is shared by subject-specific test scores cancels out. Moreover, because our vector of control variables X_{ih} includes academic year dummies, we are also controlling for potential changes in examination standards across the four years that we consider in our analysis, 2007–2010, which may cause a shift in the test scores.

Furthermore, we also compute the correction factor by using two additional methods which take account of a potential overestimation of the reliability ratio of the test scores at key stage 2. The first is identical to the method that was described above, except that the variance of the measurement error is inflated by doubling it; the second method is an approach that was suggested by Boyd *et al*. (2013), described in the on-line appendix B. It makes use of test scores observed in three different grades to derive a reliability ratio and ultimately a correction factor.

3. Institutional background

3.1. Education system in England

Full-time education in England is compulsory for all children aged between 5 and 16 years, with most children attending primary school from age 5 to 11 years and secondary school from age 11 to 16 years. 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 national tests were also carried out at key stages 1 and 3 but at present progress at these stages is examined via individual teacher assessment.

Key stage 2 National Curriculum tests are taken at the end of primary schooling, usually at age 11 years. Pupils take tests in the three core subjects of English, mathematics and science. Key stage 4 tests are taken at age 16 years at the end of compulsory schooling. Pupils enter General Certificate of Secondary Education (GCSE) or equivalent vocational or occupational examinations 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 10 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 much 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.

3.2. Exogenous variation in school funding

This section provides background on how funding was allocated to schools in the time period 2005–2010 that is considered in our empirical analysis. The aim is to show that the year-by-year variation in school resources is effectively random within schools and therefore within sibling pairs going to the same school, after controlling for observed school characteristics.

Money is allocated to schools in England from a central government schools budget using a two-stage procedure. First, central government applies a funding formula to hand out funds to 154 LEAs. These local authorities then each use their own funding formula to hand out money to schools, where funding equals expenditure. Because our analysis uses individual and sibling fixed effects estimation and we consider only siblings going to the same school, we shall not exploit between-local-authority variation in funding. For siblings within schools variation is from an increase over time in funding and slow adaptation to school level changes in educational need caused by funding rules. On average, a younger sibling in our estimation sample receives £349 more per year than her older sibling, with a standard deviation of £283. After controlling for sibling fixed effects the sibling difference in expenditure is £165 with a standard deviation of £669. Half of the siblings in our sample are two school grades apart, 20% are three grades and 30% are one grade apart.

Funding received from central government is allocated by each local authority to the schools in their area by using their own funding formulae. Apart from pupil numbers, many local authorities assess the schools' educational need according to proportions of pupils from deprived backgrounds (who are eligible for free school meals), with special educational needs and with English as an additional language in the school (Chowdry and Sibieta, 2011). When handing out funds to schools, all local authorities are, however, constrained by a minimum funding guarantee (MFG) which is set by central government. This stipulates a minimum percentage increase in funding per pupil for each school from the previous year's funding (the same across all schools in England).

In the time period that is covered by our paper about half–two-thirds of the schools budget was determined by the MFG and only the remaining budget was freely fixed by the local authority according to the schools' educational need. This implies that the funding formulae that are applied by local authorities can only partly accommodate current educational needs, with the result that schools that become more deprived from one year to the next (i.e. schools with an increased educational need) see their relative funding share falling, whereas schools experiencing a decrease in educational need see their relative funding share increase. In 2010–2011 7% of secondary schools had a level of funding at least 10% lower than predicted by using observable characteristics, and 6% had funding at least 10% higher (Chowdry and Sibieta (2011), page 12).

In our education production model we control for current school characteristics that are expected to be considered by local authorities in the funding formula and consider the remaining variation in school expenditure to be exogenous. These characteristics are based on the factors that were identified by Chowdry and Sibieta (2011), cited above, and include school size, proportion of pupils on free school meals, with first language not English, special educational needs, proportion of children from six different ethnicities and school type. In sensitivity analysis we show results of two-stage least squares estimates where we instrument school expenditure by using predicted expenditure, which is derived by adding to the lagged expenditure the percentage increase in funding per pupil set by the MFG. As we discuss in Section 5, the two-stage least squares estimates are in line with our baseline results.

Year-by-year changes in expenditure may not necessarily translate into meaningful changes in school investments, as school administrators might be reluctant to make binding decisions, such as hiring teachers, and instead spend extra funds on one-off items. However, the period 2007–2010 that is covered by our paper poses an important exception, as the MFG factor was announced in advance for a 3-year period, giving schools the security of a longer planning horizon (Sibieta, 2015). Indeed, in the four years that are covered by our empirical application, 25% of spending increases (measured as 3-year averages of the current and two preceding years) went to teachers, 24% to teaching assistants and 51% to other items. More teachers were hired and class sizes in secondary school went down by 0.7 students per class, from 16.6 students in 2007 (Sibieta, 2015). This suggests that meaningful changes in student investments have taken place in our observation period.

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 data set 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.

4.1. Outcome and observed background

Our outcomes of interest are GCSE test results at the end of compulsory schooling, usually taken at age 16 years (key stage 4). We focus on GCSEs because they mark the first major branching point in a young person's educational career. We consider key stage 4 results in the core subjects English, mathematics and science which are directly comparable with test results at the end of primary schooling. Students have the option to enter single, double or triple awards in science. These awards are designed to be of equal difficulty. Following common practices we use the best grade achieved for students entering triple science. In key stage 4 pupils receive a grade for each GCSE course based on formal examinations and some coursework elements, where pass grades include A*, A, B, C, D, E, F and G. We use a scoring system that was developed by the Qualifications and Curriculum Authority to transform these grades into a continuous point score, which we refer to as the key stage 4 score, where a pass grade G receives 16 points, and 6 points are added for each unit improvement from grade G.

We control for lagged cognitive achievement by using key stage 2 National Curriculum tests taken at the end of primary schooling, usually at age 11 years, in the three core subjects of English, mathematics and science. In the key stage 2 examinations, pupils can usually attain a maximum of 36 points in each subject, but teachers will provide opportunities for very bright pupils to test to higher levels. This practice can generate some measurement error which we consider to be random. All test scores are standardized to have a mean of 0 and a standard deviation of 1.

The NPD annual school census allows identification of a number of individual and family background variables. These include gender of the pupil, a binary variable coding ethnicity (white British, black, mixed, Indian, Pakistani or Bangladeshi and Chinese), and whether or not the first language spoken at home is English. We include in our empirical model variables indicating whether special educational needs have been identified for the child by the school or the LEA with learning difficulties, including behavioural and health conditions: those that have been assessed by LEAs receive a statement which is usually associated with additional funding received by the school; there are also pupils identified by the schools as having special needs, but without statement or whether the child has been identified by the school as being gifted and/or talented. Moreover, we can identify whether or not a pupil is eligible for free school meals. Eligibility for free school meals is linked to parents' receipt of means-tested benefits such as income support and income-based Jobseeker's Allowance and has been used in many studies as a low income marker (see Hobbs and Vignoles (2010) 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 includes only school-age siblings who are in state schools at that point in time. We also include the number of months that a pupil is older than an August-born child (the youngest in a school cohort) to control for age-at-test effects, and we use an indicator variable for the oldest pupil in a family (in the observation window 2007–2010) to control for birth order 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.

4.2. School level variables

To the NPD we merge school level expenditure information from consistent financial reporting data sets for 2004–2010. These contain details on different types of income and expenditure for each school. Assuming that pupils may benefit from school expenditure not only in their examination year, but also in the preceding years, we consider the average school expenditure over 3 years rather than yearly expenditure. 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 by using the gross domestic product deflator.

In addition we add school level characteristics to the NPD by 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 or not (community schools are owned, governed and managed by the LEA rather than by other organizations such as the Church of England in faith schools) and the number of pupils in the school (school size). We also characterize schools in terms of their pupil composition, using the proportion of pupils who receive free school meals, whose first language is English, who are of white, black, mixed, Indian, Pakistani or Bangladeshi and Chinese ethnicity and who have special educational needs. Again we average these variables describing the pupil composition over 3 years. We also add cohort mean test scores in English, science and mathematics as school level controls for prior attainment within the school, as well as academic year dummies.

4.3. Sibling definition

The NPD includes address data, which are released under special conditions, which enable us to match siblings in the data set. We have access to data from 2007, which was the first year that full address details were collected in the NPD across all pupil cohorts. Siblings are therefore defined as pupils in state schools aged 4–16 years and living together at the same address in January 2007. Siblings who are not school aged, 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 cannot distinguish them from biological siblings (see Nicoletti and Rabe (2013a) for details).

4.4. Estimation sample

For our analysis we select two samples from the NPD. The first, which we call the full sample, is a sample of students who took key stage 4 examinations in 2007 or in one of the three following years 2008, 2009 or 2010. The second sample, which we call the sibling sample, uses data for the same academic years but is restricted to siblings going to the same school. We use this sample for sibling fixed effects models and for other models when we want to compare coefficients. We exclude siblings who are in the same academic year as they do not have variation in expenditure within the same school. We keep only the oldest two siblings for each household to avoid having to expand the data set to include all sibling pair combinations within each household with the risk of overrepresenting 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 around 10000 pupils (4.5% of siblings) are third or higher order siblings in our observation window 2007–2010.

In both samples we remove pupils with duplicate data entries or with missing data on any of the background or school level variables from the data set (about 2.5% of the sample). 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%. Missing cases are concentrated among low attaining students who are more likely to be absent at the examinations or, at key stage 4, choose not to take examinations in one or more of the core subjects. Comparing the original with the retained sample the average test score is reduced by about 1%. We also exclude ‘special schools’ that exclusively cater for children with specific needs, e.g. because of physical disabilities or learning difficulties, as well as schools specifically for children with emotional and/or behavioural difficulties. Further, we exclude academy schools introduced from 2000 to allow schools more autonomy and flexible governance) for which we do not have information on expenditure, and we eliminate the top 1% in the expenditure per student distribution to avoid extreme outliers. The remaining sample contains 1 697 501 individuals of whom 339 910 are siblings as defined above. We describe the sample in the on-line appendix Table C1. To allow us to perform child fixed effects estimation across subjects we pool our data set by appending observations for test scores in English, science and mathematics for each individual. Our data set therefore contains 5 092 503 observations relating to 1 697 501 students (and 1 019 730 observations relating to 339 910 pupils for the sibling sample).

5. Empirical results

In this section we discuss our estimation results focusing on the effect of school expenditure per pupil and on the persistence ρ, which are our main coefficients of interest.

5.1. Assessing the bias caused by omitted variables

In Section 2 and the on-line appendix A we discuss the asymptotic bias that is caused by the omission of variables. In this section we evaluate the magnitude of this bias in our application by comparing the results of our two-step estimation with estimations that omit to control for some or all of the child, family and school characteristics. Specifically, we evaluate the omission biases by neglecting for the time being the measurement error issue and comparing our two-step estimation with the results from

(a) OLS estimation of the value-added model with no controls except past test scores and expenditure per pupil (OLS, no controls),
(b) OLS estimation that controls for all observed school, family and child characteristics (OLS, all controls),
(c) school fixed effect estimation, which additionally controls for unobserved school characteristics (and therefore for LEA characteristics) but not for family unobserved inputs (school fixed effects) and
(d) sibling fixed effect estimation for siblings attending the same school, which controls for both unobserved family and school characteristics, as well as for neighbourhood and LEA characteristics (sibling fixed effects).

Table 1 reports the results of the above estimation (columns (1)–(4)) and of the two-step estimation with all controls by using sibling and school fixed effects in the second step (columns (5) and (6)). The results in the top panel are based on the sample of siblings going to the same school to allow us to compare estimates across estimation methods. The results displayed in the bottom panel use the full sample and are therefore missing for models using sibling fixed effects. We compute robust standard errors by using the Huber–White estimator to allow for the possibility that the error in our model may be heteroscedastic.

Focusing first on the top panel of Table 1, column (1) displays the OLS estimates without any school, family or child controls and the estimates show a negative rather than a positive effect of per-pupil expenditure on test scores and a high persistence in cognitive skill. The negative effect of expenditure per pupil is likely to be caused by the fact that the allocation of resources

Table 1. Assessing the bias caused by omitted variables†

	Results for the following models:					
	OLS, no controls (1)	*OLS, all controls (2)*	*School fixed effects, all controls (3)*	*Sibling fixed effects, all controls (4)*	*Two-step sibling fixed effects (5)*	*Two-step school fixed effects (6)*
Sample of siblings going to same school, N = 1019730						
Expenditure per pupil	−0.040‡	−0.001	0.059‡	0.061‡	0.068‡	0.068‡
	(0.001)	(0.007)	(0.009)	(0.005)	(0.007)	(0.005)
Net persistence	0.709‡	0.583‡	0.578‡	0.503‡	0.305‡	0.305‡
	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)
Sample of all students, N = 5092503						
Expenditure per pupil	−0.038‡	−0.005‡	0.053‡	—	—	0.057‡
	(0.000)	(0.001)	(0.008)			(0.002)
Net persistence	0.713‡	0.575‡	0.570‡	—	—	0.303‡
	(0.000)	(0.000)	(0.001)			(0.001)

†Test scores are standardized. Robust standard errors (estimated by using a sandwich estimator: the Huber–White estimator) are in parentheses. Standard errors for the expenditure per pupil for the two-step estimation are bootstrapped by using 50 replications. The top panel uses the sample of siblings going to the same school; the bottom panel uses the full sample. In column (5) the net persistence is estimated by using child fixed estimation (first step), whereas the effect of expenditure is estimated using the second-step sibling fixed effect estimation. Column (6) uses school fixed effect estimation in the second step. Control variables include all variables listed in the on-line appendix Table C1 plus dummies for academic year.
‡p <0.01.

to schools is redistributive so that schools with students with more educational needs receive more money.

When we extend the model to control for all observed school, family and child characteristics, which include the characteristics that are used to determine the allocation of funds to schools from government and variables describing the school composition, the effect of expenditure per pupil on test scores is estimated to be 0 and there is a slight reduction in the persistence of the test scores (see column (2) in Table 1, top panel). The full list of control variables includes the variables in the on-line appendix Table C1 and dummies for academic year to control for possible test score inflation. Note that the sibling fixed effect estimation does not use individual level variables with no or very little variation between siblings (e.g. dummy variables for ethnic groups) because their effect would not be identified when considering differences between siblings. This estimate controlling for observed characteristics could be still biased by the omission of unobserved family, school and LEA characteristics.

Once we control for unobserved school and LEA characteristics by estimating school fixed effects, column (3), the effect of expenditure per pupil increases substantially, whereas the net persistence decreases slightly. We find that an increase in the expenditure per pupil of £1000 leads to an increase in test scores of 0.059 standard deviations, and this effect is statistically significantly different from 0 at the 1% level. Corresponding results found for English primary school pupils observed in 2001–2007 in Holmlund *et al.* (2010), who controlled for school fixed effects (but not for sibling fixed effects), are very similar (0.051, 0.040 and 0.050 standard deviations for mathematics, English and science respectively). When we additionally control for unobserved family and neighbourhood characteristics by introducing sibling fixed effects estimation for siblings going to the same school and living in the same neighbourhood, the effect of expenditure per pupil increases slightly to 0.061 and the persistence decreases to 0.503 (see column (4)).

Next we use our two-step estimation method and control for unobserved child endowments to estimate persistence, and for unobserved family, school, neighbourhood and LEA characteristics by using sibling fixed effects to estimate the expenditure effect (see column (5)). The expenditure per pupil has a marginally larger effect of 0.068 compared with sibling fixed effects that are shown in column (4), whereas the persistence decreases substantially to 0.305. Finally, in column (6) we show results for the two-step estimation using school fixed effects in the second step. The estimated expenditure effect is identical to that estimated by using sibling fixed effects, indicating that failing to control for unobserved family characteristics does not lead to a bias on this estimate.

To summarize, we find that omission of unobserved school characteristics causes a sizable bias of the school expenditure effect. Failure to account for these unobservables leads to an underestimation of the expenditure effect, whereas the estimation of the persistence in the test scores seems less affected. Controlling for family background in addition to school characteristics does not affect the results hugely. Omission of child unobserved endowment leads to a large overestimation of the net persistence, but only a modest underestimation of the expenditure effect.

Turning now to the lower panel of Table 1 which displays results based on the full sample of students in state secondary schools, we see that, apart from the OLS estimates with no controls that are displayed in column (1), all other coefficients are slightly lower than when restricting our sample to siblings in the same school. The two-step estimation with school fixed effects in the second step in the lower panel of Table 1 is our preferred estimate because it is based on the full sample of pupils rather than the subsample of siblings. Note, however, that none of the differences in the estimated effect of expenditure per pupil between the two samples are statistically significant. Moreover, if we are interested in the coefficients of other explanatory variables that are included in the education production model, we may prefer the two-step estimation with sibling fixed effects in the second step. This is because unobserved parental characteristics may differ across families within the same school and can be correlated with past test scores of the child and/or other covariates in our model. Omitting such family characteristics can cause a bias in the coefficients of explanatory variables. We provide full results for both models in the on-line appendix Table C2. The comparison shows that as expected coefficients on variables such as free-school-meal status and deprivation of neighbourhood (arguably proxies for family income) are attenuated in the two-step estimation with sibling fixed effects in the second step compared with estimates with school fixed effects in the second step. More in general, these results suggest that omitting to control for unobserved family characteristics in the second step estimation leads to an amplification bias for the effect of almost all observed child variables.

As explained in Section 2 we control for the endogeneity of school expenditure by considering a large set of school characteristics (e.g. the proportions of students from different ethnic minorities, eligible for free school meals and with special educational needs; and the average test scores at the end of primary schooling for students belongings to the same cohort). After controlling for these characteristics the remaining variation in school expenditure is related to the variation across time in the MFG, which is exogenously set at the national level.

An alternative way to exploit the exogenous variation in school expenditure is to instrument school expenditure with the minimum guaranteed expenditure, which can be computed by adding to the lagged expenditure the percentage increase in funding per pupil set by the MFG averaged across 3 years in line with the definition of expenditure in our model. This minimum guaranteed school expenditure is likely to explain actual school expenditure and is exogenous after controlling for school characteristics. Two-stage least squares estimation should provide results that are similar to our baseline results if the variation in school expenditure after controlling for school characteristics is exogenous. In the on-line appendix Table C3 we show results for two-stage least squares estimation applied to the second step of our two-step procedure. We compare the two-stage least squares estimates of the effect of school expenditure when using school fixed effects on the full sample and the sibling sample, and corresponding estimates controlling for sibling fixed effects and using the sibling sample. The estimated effects of the expenditure per pupil are all in line with the estimates that are displayed in Table 1 (point estimates are slightly higher but not statistically different from our baseline results). The endogeneity test does not reject exogeneity of the control variables for the two-stage least squares estimates by using sibling fixed effects in the second step but does reject it for the two-stage least squares estimates by using school fixed effects in the second step, suggesting that the former is the preferred estimation if we are interested in the effects of other explanatory variables.

5.2. Assessing the bias caused by measurement errors

Next we look at the role of measurement error in test scores in the estimation of the school expenditure effect. We use our two-step estimation but we correct it for measurement error bias by applying the analytic correction that was described in Section 2.2. Table 2 reports the two-step estimation without and with analytic correction (implemented by using a correction factor of 1.403) in columns (1) and (2). The measurement error seems to cause an underestimation of the net persistence but no significant differences in the effect of expenditure per pupil.

The factor analysis and the reliability ratios that we use to compute the correction factor might underestimate the variance of the measurement error var($\epsilon_{ihs,11}$) because they cannot capture measurement errors that are common across subjects and that are not related to the test specification but are related to students' characteristics such as illness during the examination period (see Boyd *et al.* (2013)). To take account of this potential underestimation we also use an alternative approach that was suggested by Boyd *et al.* (2013) and described in the on-line appendix B, which allows us to derive a correction factor of 1.550 by using observed correlations in test scores across three different grades. Furthermore, we also use a third analytic correction factor derived by doubling the variance of the measurement error. This leads to an increase in the correction factor to 1.806, which we consider as an upper bound on the true correction factor.

Estimation results by applying these second and third correction methods are reported in columns (3) and (4) of Table 2. The measurement error seems to cause an even larger underestimation of the net persistence but a non-significant difference in the effect of expenditure per pupil. This suggests that measurement errors in test scores do not lead to large biases of the effect of expenditure per pupil.

We are also concerned about the issue of potential heteroscedasticty of the measurement error. The variance of measurement errors of test scores has been found to be a U-shaped function of the ability level and to lead to a relationship between the current and lagged test score which is S shaped even if the relationship between current and lagged true ability is linear (see Boyd *et al.* (2013)). In an attempt to explore how much this heteroscedasticity issue can affect our results, we consider our child fixed effect estimation of the persistence ρ, allowing ρ to differ at the top and bottom 10th percentile of the distribution of the lagged test. We find that ρ is equal to 0.278 (standard error 0.001) at the bottom decile, 0.315 (standard error 0.001) between the 10th and 90th percentiles and 0.337 (standard error 0.001) at the top 10th percentile. These results

Table 2. Analytic correction of measurement error†

	Results for the following methods:			
	Two-step, not corrected (1)	*Two-step, analytical correction 1 (2)*	*Two-step, analytical correction 2 (3)*	*Two-step, analytical correction 3 (4)*
Expenditure per pupil	0.057‡	0.055‡	0.054‡	0.053‡
	(0.002)	(0.002)	(0.002)	(0.002)
Net persistence	0.303‡	0.425‡	0.470‡	0.547‡
	(0.001)	(0.001)	(0.001)	(0.001)
Observations	5092503	5092503	5092503	5092503

†Test scores are standardized. Bootstrapped robust standard errors are in parentheses. Control variables include all variables listed in the on-line appendix Table C1 and dummies for academic year. Analytic corrections 1, 2 and 3 are based on correction factors of 1.403, 1.550 and 1.806. The net persistence is estimated by using child fixed estimation with analytic correction (first step), whereas the effect of expenditure is estimated by using the second-step school fixed effect estimation.
‡p <0.01.

do not suggest an S-shaped relationship between the current and lagged test score; there are some statistically significant changes in the persistence across levels of the lagged test but these changes are very small. Therefore we conclude that the issue of heteroscedastic measurement errors does not seem to be a major concern in our application.

An issue that we have overlooked so far is the potential measurement error in 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 a set of alternative measures of expenditure per pupil, i.e. using the current expenditure in the key stage 4 examination year only, and using 2-, 3-, 4- and 5-year averages. Our benchmark estimation is based on a 3-year average. In Table 3 we report the results for the effect of expenditure per pupil defined by using the five definitions. In all cases we use the two-step estimation with correction for measurement error in test scores. The effect of expenditure per pupil tends to increase with the number of years that is used to compute the average expenditure per pupil but stabilizes and even decreases when using more than 4 years. This corroborates our suspicion of bigger measurement error in the yearly expenditure per pupil, which cancels out or at least reduces substantially when considering average expenditure over multiple years.

5.3. Maintained assumptions 1.1–1.6

Our value-added model imposes the following relationship between subject-specific test scores $Y_{ihs,11}$ and latent general cognitive ability $Y^*_{ih,11}$:

$$\begin{aligned} Y_{ihs,11} &= Y^*_{ih,11} + e_{ihs,11}, \\ Y_{ihs,16} &= Y^*_{ih,16} + e_{ihs,16}, \end{aligned} \tag{17}$$

Table 3. Effect of expenditure per pupil by using various measurements of expenditure†

Model	*Results for two-step school fixed effects with analytical correction*	*Observations*
Current expenditure per pupil	0.022‡ (0.001)	5092503
2-year average expenditure per pupil	0.039‡ (0.002)	5092503
3-year average expenditure per pupil	0.055‡ (0.002)	5092503
4-year average expenditure per pupil	0.073‡ (0.003)	5077644
5-year average expenditure per pupil	0.066‡ (0.003)	5037183

†Test scores are standardized. Robust standard errors (estimated by using a sandwich estimator: the Huber–White estimator) are in parentheses. Control variables include all variables listed in the on-line appendix Table C1 and dummies for academic year. The net persistence is estimated by using child fixed estimation with analytic correction (factor 1.403) and by using school fixed effects in the second step.
‡p <0.01.

where $e_{ihs,t} = v_{ihs,t} + \epsilon_{ihs,t}$ and the properties of $v_{ihs,t}$ and $\epsilon_{ihs,t}$ are described by assumptions 1.1–1.6 in Section 2. To assess the validity of these assumptions we first run an exploratory factor analysis on the three test scores at ages 11 and 16 years and find that the first factor explains on average more than 75% of the variance of the subject-specific test scores at the corresponding ages and therefore supports a single-factor model. We then estimate a structural equation model with one single factor, separately for key stages 2 and 4. Estimation results of this model are reported in Table 4. The top panel reports the factor loadings for the three subject tests where mathematics is constrained to be 1. Results suggest that the subject-specific test scores are equal to $Y^*_{ih,11}$ with factor loadings quite close to 1.

We check the assumptions that both $v_{ihs,t}$ and $\epsilon_{ihs,t}$ have equal variance across subjects and between age 11 and 16 years (see maintained assumptions 1.1 and 1.3) by looking at the variance of $e_{ihs,t} = v_{ihs,t} + \epsilon_{ihs,t}$ and find that there are statistically significant differences. Results in the second panel of Table 4 show that the percentage of total variation in subject-specific test scores explained by $e_{ihs,t}$ does indeed vary between about 20% and 30%. However, when allowing for correlation between $e_{ihs,t}$ and $e_{ihs',t}$ for $s \neq s'$ we do not reject the assumption of zero correlation in line with what is imposed by the maintained assumptions 1.2 and 1.4 (see the bottom panel in Table 4). Furthermore, we report in Table 5 correlations between test scores in mathematics, science and English at ages 11 and 16 years. The correlations are high and range from 0.611 to 0.819. We see that correlation between tests taken at two different key stages is higher when the two tests are in the same subject, and this supports maintained assumption 1.2.

We also assess whether $e_{ihs,t} = v_{ihs,t} + \epsilon_{ihs,t}$ is uncorrelated with the latent general cognitive ability $Y^*_{ih,11}$ (maintained assumption 1.5) by estimating the structural equation model for test scores at age 16 years separately for high and low ability children at age 11 years defined as pupils with an average test score across the three subjects below and above the population mean respectively. We find that the percentage of total variation in subject-specific test scores explained by $e_{ihs,t}$ varies more across subjects than across level of ability, and there does not seem to be any pattern in the relationship between the variance in measurement error and the level of ability.

Table 4. Structural equation models for the subject-specific test scores†

	Result for key stage 2 tests (1)	*Result for key stage 4 tests (2)*
	Factor loading	*Factor loading*
	Model with independent errors e_{ih_s} across s	
English	0.937‡	0.930‡
	(0.001)	(0.001)
Science	1.019‡	1.031‡
	(0.001)	(0.001)
Mathematics	1	1
$\mathrm{var}(Y^*_{ih,t})$	0.775‡	0.794‡
	(0.001)	(0.001)
Uniqueness		
$\mathrm{var}(e_{ih,\mathrm{Maths}})/\{\mathrm{var}(Y^*_{ih,t})+\mathrm{var}(e_{ih,\mathrm{Maths}})\}$	0.225	0.206
$\mathrm{var}(e_{ih,\mathrm{English}})/\{\mathrm{var}(Y^*_{ih,t})+\mathrm{var}(e_{ih,\mathrm{English}})\}$	0.292	0.283
$\mathrm{var}(e_{ih,\mathrm{Science}})/\{\mathrm{var}(Y^*_{ih,t})+\mathrm{var}(e_{ih,\mathrm{Science}})\}$	0.201	0.164
	Separate models by past high ability	*Separate models by past low ability*
$\mathrm{var}(e_{ih,\mathrm{Maths}})/\{\mathrm{var}(Y^*_{ih,t})+\mathrm{var}(e_{ih,\mathrm{Maths}})\}$	0.324	0.341
$\mathrm{var}(e_{ih,\mathrm{English}})/\{\mathrm{var}(Y^*_{ih,t})+\mathrm{var}(e_{ih,\mathrm{English}})\}$	0.469	0.395
$\mathrm{var}(e_{ih,\mathrm{Science}})/\{\mathrm{var}(Y^*_{ih,t})+\mathrm{var}(e_{ih,\mathrm{Science}})\}$	0.244	0.288
	Models allowing for correlation between errors	
$\mathrm{cov}(e_{ih,\mathrm{Maths}}, e_{ih,\mathrm{English}})$	0.000	0.000
	(3.130)	(7.162)
$\mathrm{cov}(e_{ih,\mathrm{English}}, e_{ih,\mathrm{Science}})$	0.000	0.000
	(1.868)	(5.068)
$\mathrm{cov}(e_{ih,\mathrm{Maths}}, e_{ih,\mathrm{Science}})$	0.000	0.000
	(9.479)	(2.730)

†Test scores are standardized. Results are from structural equation models assuming a single factor and constraining the factor loading for mathematics to 1.
‡$p < 0.01$.

Finally we also check whether the assumption of equal persistence for subject-specific ability (maintained assumption 1.6) is supported empirically by allowing the coefficient ρ that is estimated in our first step, child fixed effects estimation, to differ across subject (without analytical correction). We find an estimated persistence of 0.340, 0.294 and 0.276 for mathematics, English and science respectively, indicating that there are some differences but that they are not so large as to overturn our results (not reported in Table 5).

5.4. *Parametric functional form assumptions*

Because the main aim of this paper is to assess the potential biases that are caused by unobserved family, school, child, neighbourhood and LEA characteristics in linear value-added models

Table 5. Correlations between test scores

	Mathematics, key stage 4	*English, key stage 4*	*Science, key stage 4*	*Mathematics, key stage 2*	*English, key stage 2*	*Science, key stage 2*
Mathematics, key stage 4	1.000					
English, key stage 4	0.738	1.000				
Science, key stage 4	0.819	0.762	1.000			
Mathematics, key stage 2	0.767	0.612	0.674	1.000		
English, key stage 2	0.636	0.705	0.638	0.726	1.000	
Science, key stage 2	0.674	0.611	0.675	0.790	0.740	1.000

that impose constant return-to-school investments and to compare our results with previous references using such models, we work under the assumption that time invariance, additive separability and linearity hold. We test time invariance of our model by splitting our sample into two time periods (academic years 2007 and 2008, and academic years 2009 and 2010) and performing separate analysis on these samples. Our estimates of the effect of school resources are identical across the two models, and the estimated persistence is very similar. However, the assumption of additive separability and linearity are rejected. The aim of this section is to evaluate the consequences for our results and to assess the potential direction of the biases.

We assess the consequences of the linearity assumption by comparing results that were obtained by using a linear and a quadratic polynomial relationship between test scores and expenditure per pupil. Fig. 1 compares the predicted outcomes (standard deviation improvements in test scores) by using the linear and quadratic polynomial relationships and plots them against expenditure per pupil. There are some differences especially at the extremes of the expenditure distribution, but the predictions are more similar for the central part of the distribution. For this reason we think that we can interpret the effect of expenditure per pupil obtained by using the linear value-added model as an approximately unbiased effect for values of the expenditure per pupil which are not extremes. The linearity assumption is generally not rejected when using samples of smaller size than ours. Harris (2007), who used data from 32 countries with samples sizes varying between 2000 and 9000 observations, tested the assumption of linearity in school inputs and found that it cannot be rejected within countries.

To relax the assumption of additive separability we should allow for a heterogeneous effect of the expenditure per pupil by level of other inputs and by level of lagged cognitive ability (see Figlio (1999)). We cannot test whether the return-to-school expenditure changes by level of other inputs, because these are unobserved; but we can test whether the effect changes across levels of pupils' achievements measured by their past test scores at age 11 years. We do this in a companion paper using the same data (see Nicoletti and Rabe (2013b)) and we find that the effect of school expenditure is larger for children with higher test scores at age 11 years. Furthermore, we find that the estimated effect of school expenditure for students whose test scores at the end of primary schooling are close to the median is similar to the effect that was found when imposing a constant return-to-school expenditure. For this reason we expect the effect of expenditure per pupil estimated by using the linear value-added model (6) to be generally underestimated or overestimated for high or low ability children respectively, but to be approximately unbiased for pupils with a median level of test scores at age 11 years.

In conclusion, the estimated effect of expenditure per pupil when imposing linearity and additive separability is generally biased but can be considered a good proxy of the effect for

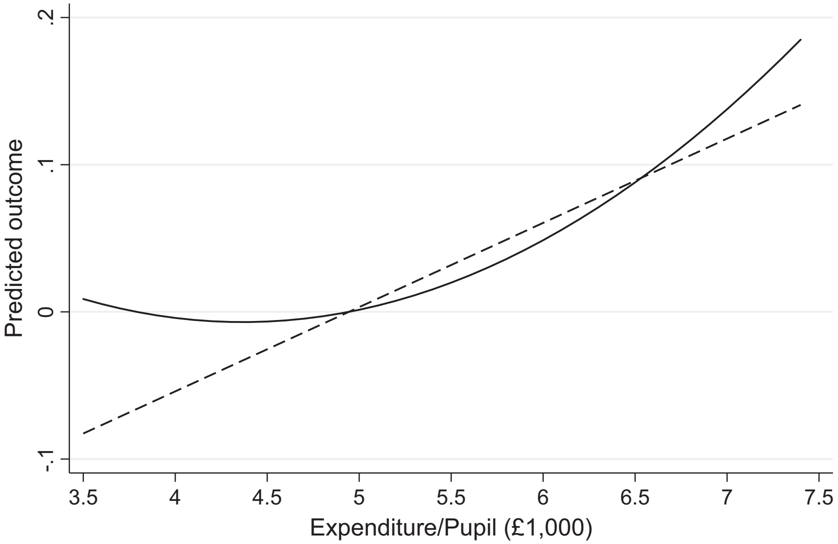


Fig. 1. Comparing linear (– – –) and quadratic (———) value-added models (NPD, 2007–2010)

levels of expenditure which are not extremes and for pupils whose lagged test scores are around the median.

6. Conclusions

In this paper, we use unusually rich English register data from the NPD to investigate biases in the estimation of the effect of school resources in value-added education production models. Econometric issues that are typically encountered when estimating education production models using administrative data are input omission and measurement error in test scores. We develop a new two-step estimation technique that tackles the endogeneity of the lagged test scores, unobserved school and family inputs and measurement error. The first step provides a consistent estimate of the persistence of achievement between education stages by applying a within-pupil, between-subject (child fixed effect) estimation which controls for the correlation between the unobserved child-specific endowment and past test scores. The second step provides a consistent estimate of the effect of school expenditure and controls for unobserved school and LEA characteristics by using school fixed effect estimation. Further, we correct for measurement error in past test scores by using analytic correction methods.

Our estimates of the effect of school spending on test scores in mathematics, English and science at the end of secondary schooling indicate that a rise in the expenditure per pupil of £1000 leads to an increase in test scores of about 6% of a standard deviation. This estimation tackles unobserved heterogeneity better than previous approaches but relies on some parametric and structural form assumptions which we discuss throughout the paper. We investigate the biases that are associated with input omission and mismeasurement by applying estimation techniques that neglect to control for some or all of the econometric issues. This enables us to assess which sources of estimation bias are most important.

To summarize, our results suggest that causal inference on the effect of school spending on student achievement requires controlling for both observed and unobserved school characteristics. The omission of such controls leads to severe underestimation of the effect. This is because schools with more disadvantaged students receive more money. However, after controlling for school differences the omission of family background does not affect the estimation of the expenditure effect. Controlling for unobserved child endowments also does not seem to bias the estimation of the spending effect. This indicates that in our quasi-experimental setting school spending is largely uncorrelated with unobserved family and child characteristics. Therefore controlling for school characteristics seems to be the main requirement to correct for the potential bias in the estimation of the school expenditure effect, supporting the credibility of previous studies based on administrative data that could not control for child and family factors to the same extent as we are (e.g. Holmlund *et al.* (2010) and Machin *et al.* (2010)). Our results do show, however, that failing to control for child endowments leads to an overestimation of the net persistence of student achievements from one stage of education to the next.

Our estimation results are important for future applications that because of data limitations are forced to estimate value-added models omitting relevant inputs and using tests with measurement error. They suggest that the most important source of bias is the omission of school characteristics, followed by the omission of family and child endowments, and lastly and least by the measurement error in the lagged test.

Acknowledgements

We thank the Department for Education for making available, under terms and conditions, data from the NPD. We also thank the Joint Editor, Associate Editor and two referees as well as participants at the European Society for Population Economics conference and the International Workshop on Applied Economics of Education for helpful comments. Financial support from the Nuffield Foundation is gratefully acknowledged. The views expressed are those of the authors and not necessarily those of the Nuffield Foundation. This work was also supported by the Economic and Social Research Council through the Research Centre on Micro-Social Change at the University of Essex (grants ES/H00811X/1 and ES/L009153/1).

References

Aitkin, M. and Longford, N. (1986) Statistical modelling issues in school effectiveness studies (with discussion). *J. R. Statist. Soc.* A, **149**, 1–42.

Altinok, N. and Kingdon, G. (2012) New evidence on class size effects: a pupil fixed effects approach. *Oxf. Bull. Econ. Statist.*, **74**, 203–234.

Altonji, J. G. and Dunn, T. A. (1996) Using siblings to estimate the effect of school quality on wages. *Rev. Econ. Statist.*, **78**, 665–671.

Andrabi, T., Das, J., Khwaja, A. I. and Zajonc, T. (2011) Do value-added estimates add value?: Accounting for learning dynamics. *Am. Econ. J. Appl. Econ.*, **3**, no. 3, 29–54.

Behrman, J., Pollak, R. and Taubman, P. (1982) Parental preferences and provision for progeny. *J. Polit. Econ.*, **90**, 52–73.

Behrman, J. R., Rosenzweig, M. R. and Taubman, P. (1996) College choice and wages: estimates using data on female twins. *Rev. Econ. Statist.*, **78**, 672–685.

Bernal, R. (2008) The effect of maternal employment and child care on children's cognitive development. *Int. Econ. Rev.*, **49**, 1173–1209.

Boardman, A. E. and Murnane, R. J. (1979) Using panel data to improve estimates of the determinants of educational achievement. *Sociol. Educ.*, **52**, 113–121.

Boyd, D., Lankford, H., Loeb, S. and Wyckoff, J. (2013) Measuring test measurement error: a general approach. *J. Educ. Behav. Statist.*, **38**, 629–663.

Chowdry, H. and Sibieta, L. (2011) School funding reform: an empirical analysis of options for a national funding formula. *Briefing Note BN123*. Institute for Fiscal Studies, London.

Clotfelter, C. T., Ladd, H. F. and Vigdor, J. L. (2010) Teacher credentials and student achievement in high school: a cross-subject analysis with student fixed effects. *J. Hum. Resour.*, **45**, 655–681.

Cunha, F. and Heckman, J. J. (2007) The technology of skill formation. *Am. Econ. Rev.*, **92**, 31–47.

Cunha, F., Heckman, J. J., Lochner, L. J. and Masterov, D. V. (2006) Interpreting the evidence on life cycle skill formation. In *Handbook of the Economics of Education* (eds E. A. Hanushek and F. Welch), ch. 12. Amsterdam: North-Hollland.

Dee, T. S. (2005) A teacher like me: does race, ethnicity, or gender matter? *Am. Econ. Rev. Pap. Proc.*, **95**, 158–165.

Dee, T. S. (2007) Teachers and the gender gaps in student achievement. *J. Hum. Resour.*, **42**, 528–554.

Department for Children, Schools and Families (2009) Departmental report 2009. *Report*. Department for Children, Schools and Families, London. (Available from `http://www.official-documents.gov.uk/document/cm75/7595/7595.pdf`.)

Department for Education (2010) Schools, pupils and their characteristics 2010. Department for Education, London. (Available from `http://www.education.gov.uk/rsgateway/DB/SFR/s000925/index.shtml`.)

Ermisch, J. and Francesconi, M. (2000) Educational choice, families, and young people's earnings. *J. Hum. Resour.*, **35**, 143–176.

Ferrão, M. E. and Goldstein, H. (2009) Adjusting for measurement error in the value added model: evidence from Portugal. *Qual. Quant.*, **43**, 951–963.

Figlio, D. N. (1999) Functional form and the estimated effects of school resources. *Econ. Educ. Rev.*, **18**, 241–252.

Frölich, M. (2008) Parametric and nonparametric regression in the presence of endogenous control variables. *Int. Statist. Rev.*, **76**, 214–227.

Fuller, W. (1986) *Measurement Error Models*. New York: Wiley.

Gibbons, S. and McNally, S. (2013) The effects of resources across school phases: a summary of recent evidence. *Discussion Paper 1226*. Centre for Economic Policy, London.

Gibbons, S., McNally, S. and Viarengo, M. (2012) Does additional spending help urban schools?: an evaluation using boundary discontinuities. *Discussion Paper 90*. London School of Economics and Political Science, London.

Goldstein, H., Rasbash, J., Yang, M., Woodhouse, G., Pan, H., Nuttall, D. and Thomas, S. (1993) A multilevel analysis of school examination results. *Oxf. Rev. Educ.*, **19**, 425–433.

Haegeland, T., Raaum, O. and Salvanes, K. G. (2012) Pennies from heaven?: Using exogenous tax variation to identify effects of school resources on pupil achievement. *Econ. Educ. Rev.*, **31**, 601–614.

Hanushek, E. A. (1979) Conceptual and empirical issues in the estimation of educational production functions. *J. Hum. Resour.*, **14**, 351–388.

Hanushek, E. A. (1986) The economics of schooling: production and efficiency in public schools. *J. Econ. Lit.*, **24**, 1141–1177.

Hanushek, E. A. (2006) School resources. In *Handbook of the Economics of Education*, vol. 2 (eds E. A. Hanushek and F. Welch), ch. 14. Amsterdam: North-Holland.

Hanushek, E. A., Rivkin, S. G. and Taylor, L. L. (1996) Aggregation and the estimated effects of school resources. *Rev. Econ. Statist.*, **78**, 611–627.

Harris, D. N. (2007) Diminishing marginal returns and the production of education: an international analysis. *Educ. Econ.*, **15**, 31–53.

He, Q., Hayes, M. and Wiliam, D. (2013) Classification accuracy in Key Stage 2 National Curriculum tests in England. *Res. Pap. Educ.*, **28**, 22–42.

Heinesen, E. (2010) Estimating class-size effects using within-school variation in subject-specific classes. *Econ. J.*, **120**, 737–760.

Hobbs, G. and Vignoles, A. (2010) Is free school meal eligibility a good proxy for family income? *Br. Educ. Res. J.*, **36**, 673–690.

Holmlund, H., McNally, S. and Viarengo, M. (2010) Does money matter for schools? *Econ. Educ. Rev.*, **29**, 1154–1164.

Jenkins, A., Levačić, R. and Vignoles, A. (2006) Estimating the relationship between school resources and pupil attainment at GCSE. *Research Report RR727*. Department for Education and Skills, London.

Krueger, A. B. (2003) Economic considerations and class size. *Econ. J.*, **113**, F34–F63.

Lavy, V. (2012) Expanding school resources and increasing time on task: effects of a policy experiment in Israel on student academic achievement and behavior. *Working Paper 18369*. National Bureau of Economic Research, Cambridge.

Lindqvist, E. and Vestman, R. (2011) The labor market returns to cognitive and noncognitive ability: evidence from the Swedish enlistment. *Am. Econ. J. Appl. Econ.*, **3**, 101–128.

Lockwood, J. R. and McCaffrey, D. F. (2014) Correcting for test score measurement error in ANCOVA models for estimating treatment effects. *J. Educ. Behav. Statist.*, **39**, 22–52.

Machin, S., McNally, S. and Meghir, C. (2010) Resources and standards in urban schools. *J. Hum. Captl*, **4**, 365–393.

Meghir, C. and Rivkin, S. G. (2011) Econometric methods for research in education. In *Handbook of the Economics of Education*, vol. 3 (eds E. A. Hanushek, S. Machin and L. Woessmann), pp. 1–87. Amsterdam: North-Holland.

Meyer, R. (1999) The production of mathematics skills in high school: What works? In *Earning and Learning: How Schools Matter* (eds S. Mayer and P. Peterson), pp. 169–204. Washington DC: Brookings Institution.

Nicoletti, C. and Rabe, B. (2013a) Inequality in pupils test scores: how much do family, sibling type and neighbourhood matter? *Economica*, **80**, 197–218.

Nicoletti, C. and Rabe, B. (2013b) School inputs and skills: complementarity and self-productivity. *Working Paper 2013-28*. Institute for Social and Economic Research, University of Essex, Colchester.

Organisation for Economic Co-operation and Development (2013) *Education at a Glance 2012: OECD Indicators*. Paris: Organisation for Economic Co-operation and Development Publishing.

Rasbash, J., Leckie, G. and Pillinger, R. (2010) Children's educational progress: partitioning family, school and area effects. *J. R. Statist. Soc.* A, **173**, 657–682.

Rivkin, S. G., Hanushek, E. A. and Kain, J. F. (2005) Teachers, schools, and academic achievement. *Econometrica*, **73**, 417–458.

Rosenzweig, M. and Wolpin, K. I. (1994) Are there increasing returns to the intergenerational production of human capital?: maternal schooling and child intellectual achievement. *J. Hum. Resour.*, **29**, 670–693.

Sass, T. R., Semykina, A. and Harris, D. N. (2014) Value-added models and the measurement of teacher productivity. *Econ. Educ. Rev.*, **38**, 9–23.

Schafer, D. W. (1986) Combining information on measurement error in the errors-in-variables model. *J. Am. Statist. Ass.*, **81**, 181–185.

Sibieta, L. (2015) The distribution of school funding and inputs in England: 1993-2013. *Working Paper W15/10*. Institute for Fiscal Studies, London.

Slater, H., Davies, N. M. and Burgess, S. (2010) Do teachers matter?: Measuring the variation in teacher effectiveness in England. *Oxf. Bull. Econ. Statist.*, **74**, 629–645.

Steele, F., Vignoles, A. and Jenkins, A. (2007) The effects of school resources on pupil attainment: a simultaneous equations multilevel modelling approach. *J. R. Statist. Soc.* A, **170**, 801–824.

Todd, P. E. and Wolpin, K. I. (2003) On the specification and estimation of the production function for cognitive achievement. *Econ. J.*, **113**, Feb., F3–F33.

Todd, P. E. and Wolpin, K. I. (2007) The production of cognitive achievement in children: home, school and racial test score gaps. *J. Hum. Captl*, **1**, 91–136.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Appendix: The effect of school spending on student achievement: addressing biases in value-added models'.