

# Accounting for Unobservables in Comparing Selective and Comprehensive Schooling \*

Stéphane Bonhomme<sup>†</sup>  
CEMFI, Madrid

Ulrich Sauder<sup>‡</sup>  
University of Warwick

August 2009

## Abstract

We compare the effects of selective and non selective secondary education on children's test scores, using British data from the National Child Development Study (NCDS). Test scores are modelled as the output of an additive production function. Inputs include family and school characteristics, as well as the child's unobserved initial endowment, which may be correlated with the education system attended. In the model, the average effect of selective education can be estimated using semiparametric Difference-in-Differences (DID) methods. We generalize the DID approach and provide conditions under which the entire counterfactual distribution of potential outcomes is identified, and can be consistently estimated using a deconvolution-related approach. Descriptive statistics on the NCDS data show that children perform better in selective schools. Our results suggest that this is essentially due to differences in pupils' composition between selective and non selective schools. When correcting for these differences, we find that the effects of selective education are small and mostly insignificant.

**JEL codes:** C33, I21.

**Keywords:** selective education, ability bias, treatment effects, quantiles.

---

\*We thank Manuel Arellano, Ghazala Azmat, Cristian Bartolucci, Olympia Bover, Martin Browning, Stefano Gagliarducci, Laura Hospido, Juan Francisco Jimeno, Per Johansson, Grégory Jolivet, Chris Taber, Ernesto Villanueva and participants at Uppsala university, Warwick University, Banco de España, RTN conference in Madrid and EEA 2007 in Budapest, and Tinbergen Institute Conference 2008 for helpful comments. We also thank Peter Shepherd, Director of Survey Operations of the Centre for Longitudinal Studies. All remaining errors are our own.

<sup>†</sup>**Corresponding author:** CEMFI, Casado del Alisal, 5, 28014 Madrid, Spain; bonhomme@cemfi.es

<sup>‡</sup>Department of Economics, Coventry CV47AL, England; usauder@gmx.net

# 1 Introduction

The dual system of secondary education that prevailed in England and Wales at the beginning of the 1970s provides an interesting case study to measure the effect of selective education on children’s outcomes. In that period, two systems coexisted, the allocation to a specific secondary school being decided at the local level by the Local Education Authority (LEA).<sup>1</sup> In the *selective* system children were assigned to two different types of secondary schools depending on their results to a test score at age 11, successful children going to “grammar” schools while the others attended less demanding “secondary modern” schools. In contrast, in the *comprehensive* system children of different ability levels were pooled together.

This coexistence was the result of an evolution. Secondary education was initially fully selective after 1945 and started to shift to comprehensive from the 1965 Crossland circular (see Crook, 2002, for a survey). As the circular did not force LEAs to switch, the shift to comprehensive was slow and heterogeneous, showing both between-LEA and within-LEA variation.<sup>2</sup> A large literature documents that the LEAs which switched first to comprehensive were different, being for example more Labour-oriented (Kerckhoff *et al.*, 1996). Also, in some LEAs secondary modern schools became comprehensive while nearby grammar schools remained selective, attracting the high ability children (Galindo-Rueda and Vignoles, 2005). As a consequence, areas where selective or comprehensive education prevailed had different compositions of pupils.

Several researchers have tried to compare the comprehensive and selective systems, controlling for differences in parental background or characteristics of the primary school (e.g., Kerckhoff, 1986, Dearden *et al.*, 2002, and Galindo-Rueda and Vignoles, 2005). These studies use data from the National Child Development Study (NCDS) and a “value-added” strategy that consists in controlling for lagged test scores in test scores equations. In a recent paper, Manning and Pischke (2006) criticize this approach. Using similar data and methods, they find a strong positive effect of attending a selective school on test scores administered at age 11, that is *before* starting secondary school. They interpret this exercise as a falsification test, which suggests that attending a selective school is

---

<sup>1</sup>In the 1960s and 1970s the UK education system was administered at the local level, and based on catchment areas (see Haydn, 2004). The LEAs had responsibility for most of the spending of secondary schools, see Chitty (2002).

<sup>2</sup>For example, in 1965 less than 5% of all public schools were comprehensive. By 1975 the proportion had reached 60%. Still today, certain LEAs such as Kent have selective (grammar) schools.

likely to be correlated with unobservables that affect later outcomes.<sup>3</sup>

In this paper, we propose a method that addresses the concern that children attending comprehensive or selective schools may have different observed *and* unobserved characteristics. Of special concern is that the child’s initial endowment, part of which reflects her cognitive ability, may be correlated with the education system attended. Because we think that our approach is original and could be applied to other settings, we devote a large part of the paper to the exposition of the methodology. Then, in a second part of the paper, we provide a substantive empirical application to the comparison of the comprehensive and selective schooling systems in England and Wales, using the NCDS data.

In the model, attending a selective secondary school is understood as a “treatment”, the effect of which we intend to measure. Test score outcomes are measured in two periods. In period 1 (age 11 in the data) the treatment is not yet realized. In period 2 (age 16 in the data) the treatment has been realized and outcomes are conditional on the education system attended. The difficulty of the exercise, in common with most of the treatment effects literature, is that we do not observe the test scores of children who attended a selective school, in the counterfactual event that they instead attended a comprehensive one. In the first part of the paper, we provide conditions under which the entire counterfactual distribution of these potential outcomes is identified.

We assume that test scores in both periods are the output of a production function (as in Todd and Wolpin, 2003, 2004). There are three types of inputs: family and school characteristics, all measured before age 11, on which we have data; the child’s initial endowment (ability), which is unobserved to the econometrician; and shocks to educational attainment, also unobserved and possibly serially correlated. The production technology that maps these three factors into test scores is additive.

We make the assumption that the fact of attending a selective or comprehensive secondary school does not depend on the shocks to future educational attainment (between ages 11 and 16). However, we allow the distribution of initial endowments of children in the two education systems to be different. For this reason, the hypothesis of selection on observables (e.g., Rosenbaum and Rubin, 1983) does not hold: if children about to attend a selective school are better endowed before starting, differences at age 16 be-

---

<sup>3</sup>The risk of finding a spurious correlation between school type and educational outcomes is not limited to the British experience. Proposed solutions involve the use of natural experiments, as in Meghir and Palme (2005) and Maurin and McNally (2006), or exploiting cross-country and time variation (Hanushek and Woessman, 2006).

tween the two education systems will not reflect the true effect of selective education on achievement, even controlling for observed covariates.

The presence of the unobserved endowment creates a challenging identification and estimation problem. We start by studying the simple setup where there are no covariates, and the returns to the endowment in period 1 (age 11) and period 2 (age 16) are equal. In this model, the average treatment effect can be consistently estimated using a Difference-in-Differences (DID) estimator. We show that the DID logic can be extended, and that the entire distribution of potential outcomes is nonparametrically identified. In particular, all quantile treatment effects are also identified. The generalization of the DID approach to the entire distribution of outcomes is based on a simple use of characteristic functions, and seems to be a new result in the literature.

Athey and Imbens (2006) also provide identification results for distributions of potential outcomes in a DID framework. However, their approach requires a *monotonicity* assumption that may be too strong in models with a linear factor structure, such as the models usually considered in the education production function literature. In our model, monotonicity only holds in the simple case where the distribution of period 2 outcomes is a mean shift of the distribution of outcomes in period 1.<sup>4</sup> In contrast, our approach allows the distributions of pre- and post-treatment outcomes to have different shapes. To our knowledge, this is the first paper to provide identification results for the entire distribution of potential outcomes in a DID setting, in the absence of monotonicity.

We then consider several extensions of the basic setup. First, we show how to deal with observed covariates, allowing the unobserved initial endowment to be correlated with covariates in an unrestricted way. So, the endowment is analogous to a “fixed” effect in a panel data model.<sup>5</sup> Next, we relax the assumption of equal returns to the endowment in the two periods. In this case, mean and distributional effects depend on the ratio of the returns in the two periods. We show that an Instrumental Variables (IV) strategy can be used to identify this ratio. Lastly, we discuss how to allow for vector-valued endowments, and how to estimate mean and distributional effects if some transformations of test scores (instead of the test scores themselves) are linear in the

---

<sup>4</sup>As in Thuysbaert (2007), who considers an additive version of the Athey and Imbens (2006) model where the distribution of the single unobservable does not change over time.

<sup>5</sup>This makes our approach different from the related work by James Heckman and coauthors, starting with Carneiro *et al.* (2003) and Hansen *et al.* (2004), where factor models are used for the unobservables. On the other hand, an important assumption in our approach is the additive index structure of the production function, which some recent work in this literature relaxes (see Cunha *et al.*, 2006).

unobserved endowment. This last extension is important, as our method is not invariant to monotone transformations of outcomes.

Estimation of mean and distributional effects is discussed next. In the case where there are no covariates, or a few discrete covariates, mean parameters can be estimated consistently in simple ways. When continuous covariates are present, we propose to use the inverse probability weighting method of Hirano, Imbens and Ridder (2003) and Abadie (2005) for estimation. To estimate the entire distribution of outcomes, we use a deconvolution estimator with trimming. The theoretical literature on nonparametric deconvolution shows that the rate of convergence of deconvolution estimators may be very slow (e.g., Carroll and Hall, 1988). However, in our application we obtained reasonably precise estimates, suggesting that the nonparametric approach that we propose is a practical possibility. We also propose a simple strategy to allow for covariates.

In the second part of the paper, we conduct an empirical analysis of the differences between selective and comprehensive education systems in the U.K. The NCDS follows children who were all born in the same week of 1958. These children were attending a secondary school between 1969 (age 11) and 1974 (age 16). Descriptive statistics show that children perform better in selective than in comprehensive schools. Moreover, regression and matching estimates controlling for a large set of family, school and local characteristics still show a positive and significant effect of attending a selective school.

Applying our methodology to the NCDS data, we find that these differences are essentially due to differences in pupils' composition. When accounting for differences in unobserved endowments as well as in observed characteristics, the mean effect becomes, in our preferred specification, small and insignificant.

We also find that quantile treatment effects are consistently zero below the median outcome. We do find some evidence of a positive effect above the median in some specifications. However, this effect is small, roughly 15% of a standard deviation at percentile 80, where the effect is maximum. Moreover, this finding is sensitive to the instruments used in order to estimate the ratio of endowments at ages 11 and 16.

Various robustness checks confirm the results. Interestingly, applying our method to test scores administered at age 11, that is before the children started secondary school, yields zero effect. This "falsification test" thus suggests that our approach controls appropriately for differences in endowments between children.

The outline of the paper is as follows. In Section 2 we present the model, motivated by our empirical application. In Section 3, we study identification in a simplified version

of the model. In Section 4, we show how to deal with observables and to allow for different returns to the endowment in the two periods. We end the methodological part of the paper in Section 5, where we discuss estimation. We then turn to the application, presenting the data in Section 6, and showing the estimation results in Section 7. Lastly, Section 8 concludes.

## 2 A model of test scores

In this section we present the model that we will use to quantify the effects of selective education.

Secondary education takes place between age 11 and age 16. We have data on test scores administered at age 11 (before secondary schooling) and at age 16. The tests are identical for all children, and are part of the NCDS survey interview. Thus they provide measures of educational achievement at both ages.

We denote as  $Y_{i1}$  a test score in period 1, that is at age 11, and as  $Y_{i2}$  a test score measured in period 2, at age 16, where  $i$  indices individual units. In the empirical part we will use the scores in mathematics at ages 11 and 16. We denote as  $D_i = 1$  (respectively  $D_i = 0$ ) the fact of attending a selective (resp. comprehensive) secondary school. We will refer to  $D_i$  as the “treatment” of interest, and try to identify and estimate its effect on children’s outcomes.

We want to compare the outcomes of the children who attended a selective secondary school with their outcomes, had they instead attended a comprehensive school. For this purpose, we need to model the counterfactual test scores that these children would have had in the comprehensive system. Following Rubin (1974) and Heckman (1990) we adopt the *potential outcome* framework and denote as  $Y_{i2}^0$  the second period outcome that individual  $i$  would have had in the absence of the treatment.  $Y_{i2}^0$  is thus the test score of individual  $i$ , had he attended a comprehensive school. Similarly, we denote as  $Y_{i2}^1$  the potential second-period outcome of  $i$ , had he attended a selective school. The observed outcome is  $Y_{i2} = D_i Y_{i2}^1 + (1 - D_i) Y_{i2}^0$ . In contrast, the outcome in period 1, before attending a secondary school, is not conditional on the education system attended between ages 11 and 16.  $Y_{i1}$  is thus a *realized*—as opposed to potential—outcome.

There are several reasons to expect  $Y_{i2}^1$  and  $Y_{i2}^0$  to be different. For example, separating children into ability groups could, by creating more homogeneous schools, facilitate the teacher’s task and improve teaching quality. Also, peer effects could play differently

in the selective and comprehensive systems (see, e.g., the discussion in Hanushek and Woessmann, 2006). Theoretical effects of selective education are discussed in the literature mentioned in the introduction. In this paper we aim at documenting the differences between  $Y_{i2}^1$  and  $Y_{i2}^0$  and do not try to identify the factors behind these differences.

We will focus on the comparison of  $Y_{i2}^1$  and  $Y_{i2}^0$  given  $D_i = 1$ . We will start with the average treatment effect on the treated (ATT):

$$\begin{aligned}\Delta &\equiv \mathbb{E}(Y_{i2}^1|D_i = 1) - \mathbb{E}(Y_{i2}^0|D_i = 1) \\ &= \mathbb{E}(Y_{i2}|D_i = 1) - \mathbb{E}(Y_{i2}^0|D_i = 1),\end{aligned}$$

where the potential outcome  $Y_{i2}^1$  for children attending a selective school coincides with the realized one.

We will also document differences in quantiles, or *quantile treatment effects* (on the treated):

$$\Delta(\tau) \equiv F_{Y_{i2}|D_i=1}^{-1}(\tau) - F_{Y_{i2}^0|D_i=1}^{-1}(\tau),$$

for any  $\tau$  between zero and one, where  $F$  is a generic notation for a cumulative distribution function (c.d.f.). Differences in quantiles are likely to be very informative in the context of selective/comprehensive education. Indeed, because of the dual nature of the selective system (separated into grammar and secondary modern schools), children at different points of the distribution could benefit differently from attending a selective school.<sup>6</sup>

We suppose the following model for potential test scores at age 16 when attending a comprehensive school ( $Y_{i2}^0$ ), and for realized test scores at age 11 ( $Y_{i1}$ ):

$$\begin{aligned}Y_{i2}^0 &= g_2^0(X_i, \eta_i, v_{i2}^0) \\ Y_{i1} &= g_1(X_i, \eta_i, v_{i1}).\end{aligned}\tag{1}$$

In (1), test scores are the output of an education production function, with three types of inputs (Todd and Wolpin, 2003, 2004). First, test scores depend on observed characteristics  $X_i$ , that include parental, school and local characteristics measured at age 11 or earlier. Importantly,  $X_i$  do not include the characteristics of the secondary school attended between ages 11 and 16. Allowing for the characteristics of the secondary school as extra inputs to the production function is difficult, due to the fact that these characteristics are influenced by the education system attended.<sup>7</sup> So, our comparison of

---

<sup>6</sup>Remark that, unlike the ATT  $\Delta$ , the quantile treatment effects  $\Delta(\tau)$  are not equal to the quantiles of the distribution of the treatment effect  $Y_{i2}^1 - Y_{i2}^0$ . As in most of the treatment effects literature, we focus on the marginal distributions of  $Y_{i2}^0$  and  $Y_{i2}^1$ . The joint distribution of  $(Y_{i2}^0, Y_{i2}^1)$  is fundamentally unidentified (Heckman, Smith and Clements, 1997), unless strong assumptions are made.

<sup>7</sup>Hence, the characteristics of the secondary school attended should be defined as ‘‘potential’’ characteristics (say,  $X_{i2}^0$  and  $X_{i2}^1$ ), had the child attended a comprehensive (or a selective) school.

the education systems will capture differences in school characteristics (such as teacher’s quality, or class size) as well as other factors (such as the fact of grouping children by ability levels).

The second input to test scores is the child’s initial endowment  $\eta_i$ , which contains the child’s cognitive ability. The initial endowment  $\eta_i$ , which influences test scores at both ages, plays a special role in the model. We allow  $\eta_i$  to be correlated with  $X_i$ , as parental inputs and school choice may be based on the child’s ability. We also allow  $\eta_i$  to be correlated with the education system attended ( $D_i$ ).

The third input to test scores are shocks to educational attainment  $v_{i1}$  and  $v_{i2}^0$ , possibly correlated with each other. Part of these shocks could reflect luck on the particular day of the exam, or an improvement or a worsening of academic achievement in a particular year, relative to the long-run academic performance of the child.

The shocks  $v_{i1}$  and  $v_{i2}^0$  will be assumed independent of  $D_i$ . This assumption is motivated by the fact that, at the time of the interviews (1969, when the pupils were 11 years old), attending a comprehensive or a selective secondary school was essentially determined by the geographic location of the family (its “catchment area”, see Haydn, 2004). We assume that location decisions are correlated with long-run characteristics of the child and the family, most importantly the academic endowment of the child, while they are uncorrelated with unexpected shocks to academic performance.<sup>8</sup> In order to make this assumption credible, we will select a sample of children who did not change secondary school. Moreover, we will make use of the many test scores that the data provide to allow for a richer structure of endowments.

Allowing for the presence of  $\eta_i$  when comparing the comprehensive and selective schooling systems is motivated by the analysis in Manning and Pischke (2006), which suggests that the distributions of unobservables in the two systems are different. However,  $\eta_i$ , which is not observed by the econometrician, acts as a “confounder” and complicates the estimation of the effects of interest  $\Delta$  and  $\Delta(\tau)$ . As the methods previously proposed in the treatment effects literature do not apply, we will need to develop new strategies for identifying and estimating these effects.

Lastly, in order to conduct the analysis we restrict the production function to be

---

<sup>8</sup>Note that, if cognitive skills (or “ability”) vary over the life of children (e.g., Cunha and Heckman, 2006), the innovations to ability between age 11 and 16 will be subsumed in  $v_{i2}^0$ . Our assumption then implies that these innovations are uncorrelated with the education system attended, only the level of ability at age 11 is. A justification could be that when parents choose where to locate (when the child is younger than 11 years old), they only observe the current cognitive skills of their child.

additive in the following sense:

$$\begin{aligned} Y_{i2}^0 &= f_2^0(X_i) + \beta_2^0 \eta_i + v_{i2}^0 \\ Y_{i1} &= f_1(X_i) + \beta_1 \eta_i + v_{i1}, \end{aligned} \tag{2}$$

where  $\beta_1$  and  $\beta_2^0$  are the scalar returns to the unobserved endowment, which may differ between the two periods.

Although it is assumed in most of the literature on the education production function, additivity may be restrictive, mostly because test scores are rather arbitrary measures of educational achievement.<sup>9</sup> In Section 4.3 we will discuss an extension to deal with transformations of the test score variables. Allowing for more general nonlinearities, and in particular relaxing the additive index structure of the model, would complicate the analysis significantly.<sup>10</sup>

In the next two sections, we will state conditions under which differences in means and quantiles of the distribution of outcomes are identified in model (2). The approach we propose is not standard, but it is related in several ways to the literature on treatment effects. To detail our method and relate it to previous work we shall first consider a simple model with no covariates and equal returns to  $\eta_i$ , and then extend the simple model to allow for observed covariates and for different returns to  $\eta_i$ .

### 3 Identification: a simplified model

We start by considering a simplified version of model (2), where there are no covariates and  $\eta_i$  has the same return in both periods ( $\beta_2^0 = \beta_1$ , normalized to 1).

#### 3.1 Model and assumptions

The model is written as follows:

$$\begin{aligned} Y_{i2}^0 &= \alpha_2^0 + \eta_i + v_{i2}^0, \\ Y_{i1} &= \alpha_1 + \eta_i + v_{i1}, \end{aligned} \tag{3}$$

where  $\alpha_1$  and  $\alpha_2^0$  are scalar parameters.

---

<sup>9</sup>In our data, test scores—which are administered in the survey interviews—are homogeneous across children in a given year. For an attempt to “anchor” test scores to a common metric (wages), see Cunha and Heckman (2006).

<sup>10</sup>The reason is that (1) can be viewed as a nonlinear panel data model with two periods. In these models, parameters of interest are often not point-identified (e.g., Honoré and Tamer, 2006). Identification of tight bounds on the parameters in non-additive models is likely to require the availability of long time-series of test scores.

We are interested in the distribution of  $Y_{i2}^0$  given  $D_i = 1$ , that is: the effect *on the treated*. In the empirical application, this is the distribution of age 16 outcomes of children who attended a selective school, in the counterfactual event that they instead attended a comprehensive school. We assume that we have data on  $(Y_{i2}, Y_{i1}, D_i)$  and study the identification of the entire counterfactual distribution of potential outcomes.<sup>11</sup>

We make three assumptions on model (3). The first one amounts to assuming that the treatment is not related to the shocks.

**Assumption 1**  $v_{i1}$  and  $v_{i2}^0$  are independent of  $D_i$ .

Assumption 1 requires statistical independence. In order to recover the mean of  $Y_{i2}^0$  given  $D_i = 1$  (and thus the average treatment effect on the treated, ATT), mean independence is sufficient. Strengthening the assumption to full independence will be useful to recover the entire distribution of  $Y_{i2}^0$  given  $D_i = 1$ , as opposed to its first moment only.

Assumption 1 is critical for identifying and estimating the effect of attending a selective school. Under Assumption 1, differences in pre-treatment outcomes reflect only differences in individual-specific characteristics  $\eta_i$ . In our application, this means that test scores at age 11 may differ on average between children who will attend a selective or a comprehensive school, but only to the extent that the average initial endowments  $\eta_i$  in the two groups are different. Likewise, the potential and realized second-period outcomes ( $Y_{i2}^0$  and  $Y_{i2}$ , respectively) may differ on average if the mean of  $\eta_i$  is not the same among treated and controls, i.e. if the initial endowments of children in the two education systems are different.

Assumption 1 implies that one can recover the mean potential outcome among the treated individuals as:

$$\mathbb{E}(Y_{i2}^0|D_i = 1) = \mathbb{E}(Y_{i2}|D_i = 0) + \mathbb{E}(Y_{i1}|D_i = 1) - \mathbb{E}(Y_{i1}|D_i = 0). \quad (4)$$

The ATT is then given by:

$$\begin{aligned} \mathbb{E}(Y_{i2}|D_i = 1) - \mathbb{E}(Y_{i2}^0|D_i = 1) &= [\mathbb{E}(Y_{i2}|D_i = 1) - \mathbb{E}(Y_{i2}|D_i = 0)] \\ &\quad - [\mathbb{E}(Y_{i1}|D_i = 1) - \mathbb{E}(Y_{i1}|D_i = 0)]. \end{aligned} \quad (5)$$

---

<sup>11</sup>In common with most of the literature on Difference-in-Differences (DID) to which our approach is related, availability of repeated cross-sections on  $(Y_{i2}, D_i)$  and  $(Y_{i1}, D_i)$  would be sufficient to apply the methods of this section. See Abadie (2005) for a discussion on the data requirements of DID with repeated cross-sections.

The right-hand side in (5) is the usual Difference-in-Differences (DID) estimand, where the additive effects of time and individual heterogeneity have been differenced out. So, DID will yield consistent estimates of the mean effect (ATT) under Assumption 1.

To recover the distribution of  $Y_{i2}^0$  given  $D_i = 1$ , we make another assumption that restricts the structure of unobservables in the model.

**Assumption 2**  $v_{i1}$  and  $v_{i2}^0$  are independent of  $\eta_i$  given  $D_i$ .

Assumption 2 requires the shocks to the outcomes to be independent of the individual-specific endowment  $\eta_i$ . This for example rules out the possibility that the shocks have individual-specific variances. Note that  $\eta_i$  is allowed to be correlated with the treatment  $D_i$ . So,  $\eta_i$  is analogous to a “fixed effect” in a panel data model, as it is independent of transitory innovations but may be correlated with  $D_i$ .<sup>12</sup> Lastly,  $v_{i1}$  and  $v_{i2}^0$  are allowed to be correlated in an unrestricted way. In the schooling application, it will be important to allow for serially correlated shocks to test scores.

Finally, we need a third, more technical assumption. Given the additivity and independence assumptions made in the model, it will be very convenient to work with *characteristic functions*. The characteristic function of a random variable  $W$  is a complex-valued function, that associates to each real number  $t$ :  $\Psi_W(t) = \mathbb{E}(\exp(jtW))$ , where  $j = \sqrt{-1}$  is a complex square root of  $-1$ .<sup>13</sup>

We will make use of three properties of characteristic functions (e.g., Lindgren, 1993, p.128-131). First, there exists a mapping between the density function<sup>14</sup> of a random variable  $W$ , say  $f_W$ , and its characteristic function  $\Psi_W$ . The link is given by the inverse Fourier transformation:

$$f_W(w) = \frac{1}{2\pi} \int \exp(-jtw) \Psi_W(t) dt. \quad (6)$$

The second important property is that, for any two *independent* random variables  $W_1$  and  $W_2$ , the characteristic function of the sum is the product of characteristic functions:  $\Psi_{W_1+W_2}(t) = \Psi_{W_1}(t)\Psi_{W_2}(t)$ .

Lastly, cumulants may be obtained using the derivatives at  $t = 0$  of the (complex) logarithm of the characteristic function, often referred to as the *cumulant generating function*:  $\kappa_W(t) = \log \Psi_W(t)$ . In particular,  $\kappa'_W(0)/j = \mathbb{E}(W)$  is the mean of  $W$ .

<sup>12</sup>In the next section,  $\eta_i$  will also be allowed to be correlated with exogenous covariates  $X_i$ .

<sup>13</sup>We use  $j$  instead of  $i$  to avoid confusion with the indexation of individual units.

<sup>14</sup>Throughout the paper, we will apply characteristic functions to continuous random variables. So, our approach does not accommodate cases where outcomes  $Y_{i1}$  or  $Y_{i2}$  are not continuously distributed.

We can now state our third assumption.

**Assumption 3** *The characteristic function of  $Y_{i1}$  given  $D_i = 0$  is non-vanishing on  $\mathbb{R}$ .*

Remark that, because of Assumptions 1 and 2, Assumption 3 is equivalent to assuming that the characteristic function of  $\eta_i$  given  $D_i = 0$  ( $\Psi_{\eta_i|D_i=0}$ ) and the characteristic function of  $v_{i1}$  ( $\Psi_{v_{i1}}$ ) have no real zeros.

It is very common in the nonparametric deconvolution literature to assume that characteristic functions have no real zeros (see Schennach, 2004, and references therein). The characteristic function of  $Y_{i1}$  given  $D_i = 0$  may have complex zeros if  $Y_{i1}$  is bounded. Real zeros arise in the case of symmetric, bounded distributions, such as the uniform. In contrast, most usual parametric distributions (normal, Gamma...) have characteristic functions with no real zeros.

### 3.2 Identification of the distribution

We now turn to the identification of the distribution of potential outcomes under the model's assumptions. To start with, the following theorem shows that the characteristic function of  $Y_{i2}^0$  given  $D_i = 1$  is identified. The proof is in Appendix A.

**Theorem 1** *Let Assumptions 1, 2 and 3 hold. Then:*

$$\Psi_{Y_{i2}^0|D_i=1}(t) = \frac{\Psi_{Y_{i1}|D_i=1}(t)}{\Psi_{Y_{i1}|D_i=0}(t)} \Psi_{Y_{i2}|D_i=0}(t). \quad (7)$$

Theorem 1 expresses the characteristic function of the potential outcome as a function of three characteristic functions that can be consistently estimated pointwise, given a random sample on  $(Y_{i2}, Y_{i1}, D_i)$ . Moreover, the theorem has an intuitive interpretation. Taking logarithms in (7) (provided they exist) we obtain:

$$\kappa_{Y_{i2}^0|D_i=1}(t) = \kappa_{Y_{i2}|D_i=0}(t) + \kappa_{Y_{i1}|D_i=1}(t) - \kappa_{Y_{i1}|D_i=0}(t). \quad (8)$$

Equation (8) is a generalization of (4) to the entire distribution. Indeed, taking first derivatives in (8) and evaluating at zero yields equation (4) for the mean effect.

In equations (4) and (8) the same logic applies: to obtain the distribution of potential outcomes, the distribution of realized outcomes in the population of treated individuals is corrected for the fact that treated and controls do not have the same distribution of unobservables  $\eta_i$ . Moreover, correcting for differences in  $\eta_i$  is done by adding and subtracting the distributional characteristics of pre-treatment outcomes for treated and

controls, respectively. This is the logic of Difference-in-Differences, that Theorem 1 extends to the entire distribution of outcomes. In the example of schooling, the age 16 test scores of children attending a selective school are thus corrected for differences in age 11 test scores between the two education systems, as children attending a comprehensive or a selective secondary school may have different initial endowments.<sup>15</sup>

Having obtained the identification of the characteristic function of potential outcomes, the identification of their density follows from the following corollary. The proof is a simple application of the inverse Fourier transformation (6).

**Corollary 1** *Let Assumptions 1, 2 and 3 hold. Then:*

$$f_{Y_{i2}^0|D_i=1}(y) = \frac{1}{2\pi} \int \exp(-jty) \left[ \frac{\Psi_{Y_{i1}|D_i=1}(t)}{\Psi_{Y_{i1}|D_i=0}(t)} \Psi_{Y_{i2}|D_i=0}(t) \right] dt. \quad (9)$$

Remark that, by integration, the cumulative distribution function (c.d.f.) of  $Y_{i2}^0$  given  $D_i = 1$ , that we denote as  $F_{Y_{i2}^0|D_i=1}$ , is also identified. Corollary 1 thus shows that the entire distribution of potential outcomes is identified. So, in addition to the ATT, the quantile treatment effects  $\Delta(\tau)$ , for  $\tau$  in  $[0, 1]$ , are also identified.

**Relationship with Athey and Imbens (2006).** Our approach is related to Athey and Imbens' (2006, AI hereafter) "Changes-in-Changes" (CIC) model. AI develop a method to estimate the entire distribution of potential outcomes in the CIC model. However, their approach requires outcomes to depend (in a nonlinear way) on a single unobserved component, the distribution of which is the same in both periods. In model (3), this assumption implies that  $v_{i1}$  and  $v_{i2}^0$  have the same distribution. Hence, under AI's assumptions, in model (3) the distribution of potential outcomes in period 2 is a simple mean shift of the distribution of realized outcomes in period 1. In the context of our application, this assumption is restrictive, as it implies that outcomes at age 11, and (potential) outcomes at age 16 in the comprehensive system, have the same dispersion and the same shape. In contrast, in our approach the distribution of  $(v_{i1}, v_{i2}^0)$  is not restricted. We think that it is important to allow for such distributional differences in our application as, for example, comprehensive schools may have an equalizing effect on children's outcomes.

---

<sup>15</sup>In addition, Theorem 1 shows that Assumptions 1, 2 and 3 have testable implications. This is so because the right-hand side in (7) must be a characteristic function. So in particular its modulus must be lower than one, as:  $|\Psi_W(t)| = |\mathbb{E}(\exp(jtW))| \leq \mathbb{E}(|\exp(jtW)|) = 1$ .

The generality of our approach is obtained at the cost of imposing additivity on the outcome variables. In particular, unlike AI our method is not invariant to monotone transformations of the test score variables. For this reason, in the empirical part we will work with various transformations of test scores.

A last specific feature of our approach relative to AI is that it uses characteristic functions instead of cumulative distribution functions (c.d.f.'s). Under AI's assumptions, the c.d.f. of potential outcomes is given by (see Theorem 3.1 in AI):

$$F_{Y_{i2}^0|D_i=1}(y) = F_{Y_{i1}|D_i=1} \left( F_{Y_{i1}|D_i=0}^{-1} \left( F_{Y_{i2}|D_i=0}(y) \right) \right).$$

The interpretation of this equation is similar to that of (7). However, AI's result applies the DID logic to c.d.f.'s, while our result applies the same logic to characteristic functions. A consequence is that, while AI obtain root- $N$  consistency for every quantile of the distribution of potential outcomes, our approach yields consistency, but at a slower rate in general (see Section 5 below for details).

## 4 Identification: the general case

In this section, we discuss the identification of mean and distributional effects in the general case, where covariates are present and returns to the endowment may vary between periods.

### 4.1 Allowing for observed covariates

The simplified model (3) requires that, in the absence of the treatment, the outcomes of the controls and the treated would have changed in the same way between periods 1 and 2. In many contexts one may want to allow for effects of covariates that are associated with the change in outcomes, and are not similarly distributed between treated and controls. This is the case in our application as, for example, children attending comprehensive schools come on average from a lower parental background, and live in less wealthy areas. In this section, we show how to extend the analysis of model (3) to allow for the presence of covariates.

Let  $X_i$  be a set of pre-treatment characteristics. In the application, examples of  $X$ 's are parental, school or local characteristics measured at age 11. Assumptions 1, 2 and 3 are assumed valid conditional on  $X_i$ . In addition, as we are interested in estimating

effects on the treated, we need the following assumption that restricts the support of the propensity score. For this we denote  $p_D = P(D_i = 1)$ , and  $p_D(x) = P(D_i = 1 | X_i = x)$ .

**Assumption 4**  $p_D > 0$ , and  $p_D(X_i) < 1$  with probability one.

Note that the assumptions restrict the correlation between transitory shocks and the treatment, yet they leave the correlation between  $\eta_i$  (and also  $v_{i1}, v_{i2}^0$ ) and  $X_i$  unrestricted. In an education production function approach, it is important not to restrict this correlation as, for example, parents may take their child's ability  $\eta_i$  into account when deciding what primary school to choose for their child (the characteristics of the primary school are part of the covariates  $X_i$ ).

Let  $\Psi_{W|Z}(t|z) = \mathbb{E}(\exp(jtW) | Z = z)$  denote the *conditional* characteristic function of a random variable  $W$  given  $Z$ . We then have the following result, that gives the identification of the conditional and unconditional characteristic functions of  $Y_{i2}^0$  given  $D_i = 1$ . The proof is in Appendix A.

**Theorem 2** *Let Assumptions 1, 2, 3 hold given  $X_i$  (almost everywhere), and let Assumption 4 hold. Then:*

$$\Psi_{Y_{i2}^0 | D_i=1, X_i}(t|x) = \frac{\Psi_{Y_{i1} | D_i=1, X_i}(t|x)}{\Psi_{Y_{i1} | D_i=0, X_i}(t|x)} \Psi_{Y_{i2} | D_i=0, X_i}(t|x), \quad (10)$$

and

$$\Psi_{Y_{i2}^0 | D_i=1}(t) = \frac{1}{p_D} \mathbb{E}[\omega(t|X_i) (1 - D_i) \exp(jtY_{i2})], \quad (11)$$

where we have denoted as

$$\begin{aligned} \omega(t|X_i) &\equiv \frac{p_D(X_i)}{(1 - p_D(X_i))} \frac{\Psi_{Y_{i1} | D_i=1, X_i}(t|X_i)}{\Psi_{Y_{i1} | D_i=0, X_i}(t|X_i)} \\ &= \frac{\mathbb{E}[D_i \exp(jtY_{i1}) | X_i]}{\mathbb{E}[(1 - D_i) \exp(jtY_{i1}) | X_i]}. \end{aligned} \quad (12)$$

As in the case without covariates, knowledge of the characteristic function implies knowledge of the density, using the inverse Fourier transformation (6).

**Comparison with selection on observables.** It is interesting to compare these identification results to the literature on selection on observables and estimation using matching. Under the selection on observables assumption (e.g., Rosenbaum and Rubin, 1983), potential outcomes are independent of the treatment given observables. In our notation, this means that  $Y_{i2}^0$  is independent of  $D_i$  given  $X_i$ , hence  $f_{Y_{i2}^0 | D_i=1, X_i} = f_{Y_{i2} | D_i=0, X_i}$ .

One may use this identity of conditional densities (and also c.d.f.'s) to derive results on unconditional quantiles of the distribution of potential outcomes, as in Firpo (2007).

However, in model (3) selection on observables does not hold. Instead, the model satisfies an assumption of selection on observables *and unobservables*, as  $Y_{i2}^0$  is independent of  $D_i$  given  $X_i$  and  $\eta_i$ . As the distribution of  $\eta_i$  may differ between treated and controls, estimators based on the selection-on-observables assumptions will be biased. The empirical part will illustrate that taking into account differences in unobservables may be very important in schooling applications.

## 4.2 Allowing for different returns to unobservables

The benchmark model (3) imposes that the coefficients of  $\eta_i$  in the equations of pre- and post-treatment outcomes are the same. In some instances, one may want to allow for different coefficients. This is the case in the schooling application, where ability may be differently rewarded at age 11 and 16, and may have a specific return in the comprehensive education system. Here we show how to extend our framework to allow for different coefficients in both periods.

We start with the following model, without observed covariates:

$$\begin{aligned} Y_{i2}^0 &= \alpha_2^0 + \beta_2^0 \eta_i + v_{i2}^0, \\ Y_{i1} &= \alpha_1 + \beta_1 \eta_i + v_{i1}, \end{aligned} \tag{13}$$

where  $\alpha_1$ ,  $\beta_1$ ,  $\alpha_2^0$  and  $\beta_2^0$  are scalar parameters.

Note that (13) implies that

$$Y_{i2}^0 = \alpha_2^0 - \rho \alpha_1 + \rho Y_{i1} + v_{i2}^0 - \rho v_{i1}, \tag{14}$$

where  $\rho = \beta_2^0 / \beta_1$  is the ratio of returns to  $\eta_i$ .

$Y_{i1}$  is endogeneous in equation (14), if only because of the presence of the contemporaneous shock  $v_{i1}$ . In similar contexts, solutions often involve the use of instrumental variables. The next identifying assumption requires that such an instrument is available.

**Assumption 5** *There exists a variable  $\tilde{Y}_{i0}$  such that:*

$$\begin{cases} v_{i1} \text{ and } v_{i2}^0 \text{ are uncorrelated with } \tilde{Y}_{i0} \text{ given } D_i = 0, \\ Y_{i1} \text{ and } \tilde{Y}_{i0} \text{ are correlated given } D_i = 0. \end{cases} \tag{15}$$

Under Assumption 5,  $\tilde{Y}_{i0}$  is a valid instrument for  $Y_{i1}$  in (14) when conditioning on  $D_i = 0$ . So  $\rho$  is identified as:

$$\rho = \frac{\text{Cov}\left(\tilde{Y}_{i0}, Y_{i2} | D_i = 0\right)}{\text{Cov}\left(\tilde{Y}_{i0}, Y_{i1} | D_i = 0\right)}. \quad (16)$$

In the application to schooling, we will use lagged test scores to instrument  $Y_{i1}$  in (14). Lagged dependent variables are often used as instruments in linear panel data models (e.g., Holz-Eakin *et al.*, 1988). However, lagged test scores in mathematics will be invalid instruments in general if the shocks to test scores are serially correlated at all lags. For this reason, we will contrast the results obtained using various sets of instruments, namely using lagged scores in other subjects, and using father's and mother's education as instruments.

As an alternative strategy, we will also allow for various (two or three) unobserved endowments. One reason for allowing for various  $\eta$ 's in the application is that there is strong evidence, in psychology and in economics, that cognitive ability is multidimensional (see, e.g., Heckman *et al.*, 2006). Another reason is that some shocks to test scores may be very persistent, and may not be adequately controlled for by one single initial endowment. Our framework can easily be generalized to allow for a vector of endowments, provided that one has data on various pre-treatment outcomes (e.g., several tests administered at age 11) and various instruments. Indeed, in that case,  $\beta_2^0$  is a row vector,  $\beta_1$  is a matrix, and  $\rho = \beta_2^0 \beta_1^{-1}$  is a two or three-dimensional parameter row vector.  $\rho$  is then identified from an equation similar to (16) where the scalar covariances are replaced by vectors/matrices. Note that the various components of the vector  $\eta_i$  are allowed to be correlated with each other in an unrestricted way.

Provided that  $\rho$  is identified, we may then rewrite (13) as:

$$\begin{aligned} Y_{i2}^0 &= \alpha_2^0 + \beta_2^0 \eta_i + v_{i2}^0, \\ \rho Y_{i1} &= \rho \alpha_1 + \beta_2^0 \eta_i + \rho v_{i1}. \end{aligned} \quad (17)$$

Note that (17) is still valid when the unobserved endowment  $\eta_i$  is a column vector (as well as  $Y_{i1}$ ), and  $\rho$  and  $\beta_2^0$  are row vectors.

Model (17) is formally equivalent to the benchmark model (3) with equal returns in the two periods, so the analysis of Section 3 can be replicated, under the same assumptions. In particular, Theorem 1 implies that, if Assumptions 1, 2, 3 and 5 hold, then the characteristic function of potential outcomes is identified as:

$$\Psi_{Y_{i2}^0 | D_i=1}(t) = \frac{\Psi_{Y_{i1} | D_i=1}(\rho t)}{\Psi_{Y_{i1} | D_i=0}(\rho t)} \Psi_{Y_{i2} | D_i=0}(t). \quad (18)$$

In logarithms, we obtain:

$$\kappa_{Y_{i2}^0|D_i=1}(t) = \kappa_{Y_{i2}|D_i=0}(t) + \kappa_{Y_{i1}|\widetilde{D}_i=1}(\rho t) - \kappa_{Y_{i1}|D_i=0}(\rho t). \quad (19)$$

The density of potential outcomes is thus identified, as:

$$f_{Y_{i2}^0|D_i=1}(y) = \frac{1}{2\pi} \int \exp(-jty) \left[ \frac{\Psi_{Y_{i1}|D_i=1}(\rho t)}{\Psi_{Y_{i1}|D_i=0}(\rho t)} \Psi_{Y_{i2}|D_i=0}(t) \right] dt. \quad (20)$$

Hence the identification of the mean and quantiles of that distribution.

We can similarly prove the identification in the model with observables  $X_i$ , when Assumption 5 holds given  $X_i$ . In particular,  $\widetilde{Y}_{i0}$  may be correlated with  $X_i$ . Lastly, note that  $\rho$  may depend on  $X_i$ . In practice, we found that imposing that  $\rho$  is constant improved the precision of the density estimates. We shall make this assumption in the application.

**Comparison with the value-added methodology.** The “value-added” methodology<sup>16</sup> is widely used to estimate test score equations. However, it yields an inconsistent estimate of the mean of potential outcomes in model (13). To see why, remark that taking expectations in (14) yields, using Assumption 1:

$$\mathbb{E}(Y_{i2}^0|D_i = 1) = \alpha_2^0 - \rho\alpha_1 + \rho\mathbb{E}(Y_{i1}|D_i = 1).$$

The value-added estimand of the mean, in contrast, is:

$$\begin{aligned} \mathbb{E}(\mathbb{E}(Y_{i2}|Y_{i1}, D_i = 0) | D_i = 1) &= \alpha_2^0 - \rho\alpha_1 + \rho\mathbb{E}(Y_{i1}|D_i = 1) \\ &\quad + \mathbb{E}(\mathbb{E}(v_{i2}^0 - \rho v_{i1} | Y_{i1}, D_i = 0) | D_i = 1) \\ &= \mathbb{E}(Y_{i2}^0 | D_i = 1) \\ &\quad + \mathbb{E}(\mathbb{E}(v_{i2}^0 - \rho v_{i1} | Y_{i1}, D_i = 0) | D_i = 1). \end{aligned} \quad (21)$$

Equation (21) shows that the value-added estimand is *not* equal to the mean of potential outcomes. As a consequence, the value-added estimand of the ATT is not equal to the true ATT. To gain intuition, consider the special case where the shocks  $v_{i1}$  and  $v_{i2}^0$  are uncorrelated and  $\rho$  is positive. Then the value-added estimand in (21) is an *underestimate* of the mean. So the value-added estimand of the ATT is an *overestimate* of the ATT. This suggests that estimates of the effect of selective schooling based on the value-added methodology may overestimate the true effect. The magnitude of the overestimation will depend on  $\rho$ , and also on the variance of the shock in period 1 ( $v_{i1}$ ) relative to that of  $Y_{i1}$ .

---

<sup>16</sup>We refer to the “value-added methodology” as the approach that consists in controlling (not necessarily linearly) for lagged test scores in test score equations.

### 4.3 Nonlinearities in the production function

The assumption that the education production function is linear in the endowment may be too strong. In particular, unlike the approach in Athey and Imbens (2006), ours is not invariant to monotone transformations of the test scores. So it is important to check the robustness of our results to other normalizations of the scores.

To simplify the presentation, consider a model without covariates, where the dependent variables are now some transformations of the original test scores:

$$\begin{aligned} h(Y_{i2}^0; \lambda_2^0) &= \alpha_2^0 + \beta_2^0 \eta_i + v_{i2}^0, \\ h(Y_{i1}; \lambda_1) &= \alpha_1 + \beta_1 \eta_i + v_{i1}, \end{aligned} \tag{22}$$

where  $\alpha_1$ ,  $\beta_1$ ,  $\alpha_2^0$  and  $\beta_2^0$  are scalar parameters,  $\lambda_1$  and  $\lambda_2^0$  are vectors of parameters, and  $h$  is a known function that we suppose increasing in  $y$ . One example, widely used in practice, is the Box-Cox transformation:

$$h(y; \lambda) = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \text{if } \lambda > 0, \\ \log y & \text{if } \lambda = 0. \end{cases} \tag{23}$$

Remark that, although model (22) is not additive anymore, it has an additive index structure. Relaxing this assumption would require the use of very different methods for identification and estimation.

Suppose to start with that  $\lambda_1$  and  $\lambda_2^0$  are known. Then the methods exposed in the previous sections imply, under the aforementioned assumptions, that the distribution of  $h(Y_{i2}^0; \lambda_2^0)$  given  $D_i = 1$  is identified. It follows from basic statistical theory that the distribution of  $Y_{i2}^0$  given  $D_i = 1$  is also identified. For example, the correspondence between the quantiles of the two distributions is given by:

$$F_{Y_{i2}^0|D_i=1}^{-1}(\tau) = h^{-1} \left[ F_{h(Y_{i2}^0; \lambda_2^0)|D_i=1}^{-1}(\tau); \lambda_2^0 \right],$$

where  $h^{-1}(h(y; \lambda); \lambda) = y$ .

Moreover, knowledge of the distribution of potential outcomes implies that one can recover the mean of  $Y_{i2}^0$  given  $D_i = 1$ , as:

$$\mathbb{E}(Y_{i2}^0|D_i = 1) = \int y f_{Y_{i2}^0|D_i=1}(y) dy.$$

This would not be possible if only the mean of the transformed outcome was identifiable, as opposed to its full distribution. See Abrevaya (2002) for a related point in the context of a Box-Cox transformation model.

Finally, to study the identification of  $\lambda_1$  and  $\lambda_2^0$ , a possibility is to take quasi-differences in (22), yielding:

$$h(Y_{i2}^0; \lambda_2^0) = \alpha_2^0 - \rho\alpha_1 + \rho h(Y_{i1}; \lambda_1) + v_{i2}^0 - \rho v_{i1}, \quad (24)$$

where, as before,  $\rho = \beta_2^0/\beta_1$ . Using sufficiently many instruments  $\tilde{Y}_{i0}$  (e.g., lagged test scores), it is then possible to jointly identify all the parameters in (24), using nonlinear IV (e.g., Amemiya, 1985, p.250).

## 5 Estimation

In this section we discuss estimation of mean and distributional effects. A STATA program is available online.<sup>17</sup>

### 5.1 No covariates

We start with the simple model (3) without covariates, where the coefficients of  $\eta_i$  are the same in both periods. Throughout, we will assume that we have a random sample  $(Y_{i2}, Y_{i1}, D_i)$ ,  $i = 1, \dots, N$ .

Estimating the mean in (4) is straightforward, so we concentrate on the estimation of the density of potential outcomes in (9). Our pointwise estimate is:

$$\hat{f}_{Y_{i2}^0|D_i=1}(y) = \frac{1}{2\pi} \int_{-T_N}^{T_N} \exp(-jty) \left[ \frac{\hat{\Psi}_{Y_{i1}|D_i=1}(t)}{\hat{\Psi}_{Y_{i1}|D_i=0}(t)} \hat{\Psi}_{Y_{i2}|D_i=0}(t) \right] dt. \quad (25)$$

In this equation,  $\hat{\Psi}_W$  denotes the *empirical characteristic function* of  $W$ . For example, if  $N_0$  denotes the number of individuals in the control group (comprehensive):

$$\hat{\Psi}_{Y_{i2}|D_i=0}(t) = \frac{1}{N_0} \sum_{i, D_i=0} \exp(jtY_{i2}).$$

$T_N$  is a trimming parameter that ensures that the integral in (25) is finite. To guarantee the consistency of the estimator,  $T_N$  needs to tend to infinity with the sample size  $N$ .<sup>18</sup>

The density estimator given by (25) is very similar in spirit to deconvolution estimators considered, for example, in Carroll and Hall (1988) and Fan (1991). It is out of

<sup>17</sup>See <http://www.cemfi.es/~bonhomme/>

<sup>18</sup>Note that (25) does not ensure that  $\hat{f}_{Y_{i2}^0|D_i=1}$  is a valid density. In practice, we imposed that the estimated density is positive (by taking the positive value) and that it integrates to one (by simple rescaling). We also rescaled the characteristic function so that the mean of the distribution coincides with the mean estimated using an empirical counterpart of (4). These operations had very little effect on the estimated densities and quantiles.

the scope of this paper to study its asymptotic properties. However, in analogy with the deconvolution literature we conjecture that the estimator is consistent, and that its rate of convergence depends on the tails of the characteristic functions that appear in (25). In particular, it is likely that the estimator  $\hat{f}_{Y_{i2}|D_i=1}(y)$  (for a given  $y$ ) is *not* root- $N$  consistent. The same is likely to hold for quantiles  $\hat{F}_{Y_{i2}|D_i=1}^{-1}(\tau)$ , for  $\tau$  between zero and one. Slow rates of convergence are the price to pay for estimating distributions in a model where neither parametric nor monotonicity assumptions (Athey and Imbens, 2006) are made.

To restore root- $N$  consistency, one possibility would be to adopt a flexible parametric specification for the distributions. Compared to this alternative, our approach is non-parametric, and does not require to specify the dependence of  $D_i$  on  $\eta_i$  (and  $X_i$ , when observables are present). Moreover, the empirical results that we obtain are sufficiently precise to be informative in many cases. So the slow convergence rate of the estimator does not seem to be a severe problem in the present analysis.

Lastly, extending the estimator in (25) to cases where the returns to  $\eta_i$  are different in the two periods is almost immediate. We propose to proceed in two steps. First, we estimate  $\rho$  by an IV regression of  $Y_{i2}$  on  $Y_{i1}$  on the subsample of observations with  $D_i = 0$ , using  $\tilde{Y}_{i0}$  as an instrument. This yields  $\hat{\rho}$ . In a second step, the density is estimated as:

$$\hat{f}_{Y_{i2}|D_i=1}(y) = \frac{1}{2\pi} \int_{-T_N}^{T_N} \exp(-jty) \left[ \frac{\hat{\Psi}_{Y_{i1}|D_i=1}(\hat{\rho}t)}{\hat{\Psi}_{Y_{i1}|D_i=0}(\hat{\rho}t)} \hat{\Psi}_{Y_{i2}|D_i=0}(t) \right] dt. \quad (26)$$

We use a numerical approximation to compute the integral in (26).<sup>19</sup> Then, once the density is estimated, the c.d.f. is directly obtained by numerical integration. Lastly, quantiles are computed by inversion of the estimated c.d.f.

## 5.2 Estimation in the presence of covariates

The previous analysis can be easily extended to allow for covariates  $X_i$ , in cases where there are only few discrete covariates. In our application, we want to condition on many covariates, discrete and continuous, such as parental and school characteristics. We proceed as follows.

In order to estimate  $\rho$ , we regress  $Y_{i2}$  on  $Y_{i1}$  and  $X_i$  by 2-Stage Least Squares (2SLS) on the subsample of observations such that  $D_i = 0$ , using  $X_i$  and  $\tilde{Y}_{i0}$  as instruments.

---

<sup>19</sup>We use the trapezoid rule, with 200 equidistant nodes.

Then the ATT satisfies:

$$\begin{aligned} \mathbb{E}(Y_{i2}|D_i = 1) - \mathbb{E}(Y_{i2}^0|D_i = 1) &= \mathbb{E}(Y_{i2} - \rho Y_{i1}|D_i = 1) \\ &\quad - \mathbb{E}(\mathbb{E}(Y_{i2} - \rho Y_{i1}|D_i = 0, X_i)|D_i = 1). \end{aligned} \quad (27)$$

Following Hirano *et al.* (2003) and Abadie (2005), (27) can be shown to be equivalent to

$$\mathbb{E}(Y_{i2}|D_i = 1) - \mathbb{E}(Y_{i2}^0|D_i = 1) = \frac{1}{p_D} \mathbb{E} \left\{ \frac{p_D(X_i)}{1 - p_D(X_i)} (D_i - p_D(X_i)) (Y_{i2} - \rho Y_{i1}) \right\}. \quad (28)$$

We estimate the unconditional ATT as an empirical analog of (28). Namely, we estimate the propensity score  $p_D(X_i)$  by logit, which may be viewed as a first (parametric) approximation to the series logit estimator used in Hirano *et al.* (2003). We also replace  $\rho$  in (28) by  $\hat{\rho}$ .<sup>20</sup> Note also that, for (28) to be well-defined, we need the propensity score to be strictly lower than 1. In practice, we selected the observations for which the propensity score is between its 5<sup>th</sup> and 95<sup>th</sup> percentiles. Finally, for comparison purposes, in the empirical analysis we will also report ATT estimates obtained by replacing expectations by linear projections in (27).

We estimate the counterfactual density of outcomes as:

$$\hat{f}_{Y_{i2}^0|D_i=1}(y) = \frac{1}{2\pi} \int_{-T_N}^{T_N} \exp(-jty) \frac{1}{\hat{p}_D} \left( \frac{1}{N} \sum_{i=1}^N \hat{\omega}(\hat{\rho}t|X_i) (1 - D_i) \exp(jtY_{i2}) \right) dt, \quad (29)$$

where  $\hat{\omega}(t|X_i)$  is an estimate of  $\omega(t|X_i)$  given by (12). To compute  $\hat{\omega}(t|X_i)$  for all  $t$ <sup>21</sup> and covariates values, we replace the conditional expectations in (12) by linear projections. As before, to compute  $\hat{f}_{Y_{i2}^0|D_i=1}$  we use numerical integration. Finally, to choose the trimming parameter  $T_N$ , we use a simple method due to Diggle and Hall (1993). See Appendix B for more details. C.d.f.'s and quantiles are then estimated as in the case without covariates.

Our density estimator thus relies on linearizing the conditional expectations. This imposes more structure, in order to deal with the curse of dimensionality created by the presence of many regressors. However, our experiments which consisted in adding squares and interactions as extra regressors in  $X_i$  yielded very similar estimation results, suggesting that the linearization is a reasonable approach in this case (see Section 7).

---

<sup>20</sup>In the empirical part we will also report *matching* estimates, that account for selection on observables only. In that case, we will set  $\rho$  to zero in (28).

<sup>21</sup>In practice, a grid of 200 points  $t$  was used.

## 6 Data and descriptive evidence

In this section we present the NCDS data. The next section will present the estimation results based on the framework described in the first part of the paper.

**Sample selection.** The NCDS is an ongoing longitudinal survey of a British birth cohort born between March 3 and 9 of 1958. The initial sample consisted of 17,634 individuals which were resurveyed on seven further occasions in order to monitor their changing health, education, social and economic circumstances. Attrition has led the sample size to shrink in the subsequent waves:<sup>22</sup> in 1965 (when the children were 7 years old), 1969 (age 11), 1974 (age 16), 1981, 1991, 1999/2000, and 2004 (when the cohort had reached the age of 46).

We are interested in the effects of selective and comprehensive schooling on outcomes. For this reason, we exclude from our sample other types of schools, such as technical or private schools. In addition, we only keep children who attended the same school during their five years of secondary schooling.

Private schools represent a small percentage of all schools (6.7% of the observations in the original NCDS data). However, private schools may have been a way for parents to escape the comprehensive system. The same argument could apply to mobility between comprehensive and selective schools. We checked, using the full NCDS data, that the way we restricted the sample is unlikely to bias the results. First, we found no evidence of an association between the share of comprehensive schools in the LEA and the probability of attending a private school, conditional on parental characteristics. Second, although the extent of between-LEA mobility is not negligible (11% of children changed LEA between age 11 and age 16), we found that the share of comprehensive schools in the LEA of origin had little influence on the probability of moving to another LEA.<sup>23</sup> Appendix D gives more details. Lastly, note that our comparison of selective and comprehensive schooling will not be biased, if the decision to enter a private school or move to another school is based on the observed characteristics that we control for, and the unobserved endowment.

---

<sup>22</sup>We will treat attrition as exogenous, consistently with most previous studies using this dataset. Attrition is rather severe in the NCDS. Moreover, Connolly *et al.* (1992) show that attrition is somewhat stronger among children from lower parental background. So the results of this paper might not be representative of the full NCDS cohort.

<sup>23</sup>Documenting between-LEA mobility is useful as, in the period we study, pupils were in general allocated to a specific secondary school by the LEA (Haydn, 2004). So, mobility between comprehensive and selective schools occurred mostly as the consequence of mobility between LEAs.

An important note of caution concerning the data is that the definitions of “comprehensive” and “selective” schools that we use refer to the status of the school in 1974. There is no information in the NCDS on the characteristics of the secondary school in 1969. So, it may be that a child entered a selective school at 11, that her school became comprehensive before she completed her secondary education, and thus that she is classified as attending a “comprehensive” school. Hence part of the effect we measure will reflect the effect of the comprehensivisation reform on students’ performance. In Section 7 we will provide different estimates, depending on the date when the school became comprehensive.

**Choice of variables.** We obtain a sample of 6870 observations, 56% of the children attending a comprehensive school. Our measure of children’s outcomes is the test score in mathematics administered at age 16. As other test scores, it was given during the survey interview.<sup>24</sup> We also use test score variables measured before starting secondary school. These are test scores in mathematics and reading administered at ages 7 and 11, a verbal test score administered at age 11, and two additional tests administered at age 7: “draw-a-man” and “copying designs”. More details about those tests are given in Appendix C.

The control variables that we use can be divided into three categories. Family characteristics include the gender of the child, father’s and mother’s education, the father’s social class, father’s and mother’s income (both reported in brackets), and the labor market status of the mother. School attributes include pupil-teacher ratios at ages 7 and 11, the nature of the primary school and the existence of ability tracking. We do not include 1974 school characteristics as controls. Local characteristics contain percentages of unemployed workers in the ward where the child lives and other percentages that we constructed by merging the NCDS with census data for 1971. Lastly, in some specifications we included the share of comprehensive schools in the LEA as an additional control, as well as additional LEA characteristics. References about the data sources and the variables used are given in Appendix C.

**Descriptive statistics.** Table 1 shows some descriptive statistics for the two groups of children in the sample, attending the comprehensive or the selective education sys-

---

<sup>24</sup>We do not use the reading score administered at age 16 in this paper. Reading questionnaires at ages 11 and 16 were identical, and the age 16 scores are concentrated at high values and show little variation above the median.

tem. Children aged 16 attending selective schools score on average 1.9 points higher in mathematics than children attending comprehensive schools, roughly 30% of a standard deviation. The table shows that the two systems are also very different in terms of intake, as children attending selective schools perform better at all tests at ages 7 and 11. For example, they score 3 points higher in mathematics at age 11, that is 30% of a standard deviation. We can also observe that the standard deviation of the age 11 score in mathematics is larger for children attending a school in the selective system later on. Lastly, the parents of children attending selective schools are also slightly more educated.

The selective system is by construction very heterogeneous. Table 2 illustrates this feature, showing descriptive statistics by school type: grammar and secondary modern. The table shows huge differences, children in grammar schools scoring 10 points more than the ones in secondary modern schools. Moreover, there are also marked differences in terms of intake. For example, children at grammar schools score on average 15 points higher in mathematics at age 11, and their parents are more educated.

The strong correlations between age 11 test scores and the type of secondary school attended suggests that the pupils' composition of comprehensive and selective (grammar or secondary modern) schools is very different. In the next section, we apply our methodology to compare the two schooling systems, allowing for differences in observable and unobservable characteristics between pupils.

## 7 Estimation results

### 7.1 Mean effects

We start by documenting the mean effect of attending a selective school, on the treated, i.e. for children who actually attended a selective school.

The first step in the estimation of the mean effects is to estimate the ratio of the returns to the unobserved endowment  $\eta_i$  between age 11 and age 16,  $\rho = \beta_2^0/\beta_1$ . We estimate  $\rho$  by a 2-Stage Least Squares (2SLS) regression of the score in mathematics at age 16 on the score in mathematics at age 11 and exogenous covariates, on the subsample of children attending a comprehensive school ( $D_i = 0$ ). We contrast various sets of instruments. We start by using lagged scores (administered at age 7) in mathematics and reading. Lagged dependent variables are often used as instruments in linear panel data models. However, if the shocks to test scores at age 7 and 11 are correlated, those instruments will be invalid in general. For this reason, we also present results using as

instruments test scores administered at age 7 in other subjects: draw-a-man and copying designs (see Section 6), which are less correlated with the scores in mathematics.<sup>25</sup> Lastly, we also show results using father’s and mother’s education as instruments. Parental education will be a valid instrument for  $Y_{i1}$  if it is a determinant of the child’s endowment, uncorrelated with future shocks to educational attainment.<sup>26</sup> Table 3 presents the results, for various specifications of covariates.

The  $\rho$  estimates vary little with the set of covariates. However, they depend rather strongly on the set of instruments used. The ratio of returns increases from 0.52 to 0.56 when draw-a-man and copying designs scores are used as instruments instead of mathematics and reading, and increases to 0.68 when parental education is used instead. Note that in this latter case, the instruments are weaker (low partial  $R^2$ ) and the precision of the estimates is lower. Because of the variation in the  $\rho$  estimates across specifications, in the rest of this section we will present the results for each of the three sets of instruments.

We then turn to the estimates of the average treatment effects on the treated (ATT)  $\Delta$ . The four first rows in Table 4 show the effects obtained when accounting for selection on observables only. Rows 1 and 3 present the estimates of the coefficient of the dummy variable of attending a selective secondary school in the regression of the test score in mathematics at age 16, controlling for exogenous covariates (row 1) and for exogenous covariates and the test score in mathematics at age 11 (row 3). Rows 2 and 4 present the corresponding matching estimates, using the inverse probability weighting method of Hirano *et al.* (2003).

In our favorite covariates specification (3), which includes family, school and local controls, the regression and matching estimates are about 1.5 points in mathematics. This effect drops to roughly 0.4 points when including the lagged test score as a control. This represents a modest, but positive and significant, gain associated with attending a selective school, 7% of a standard deviation. The matching estimate including the lagged maths score as a control is of similar magnitude, though insignificant.

Rows 5 to 10 in Table 4 show the ATT estimates, when accounting for differences in observables and unobservables between the two education systems. Rows 8 to 10 show matching estimates, estimated using an empirical counterpart of (28). Rows 5 to 7 show

---

<sup>25</sup>See Table A5 in the appendix for the sample correlations between test scores.

<sup>26</sup>When using father’s and mother’s education as instruments, we drop these variables from the set of exogenous covariates. However, the family characteristics included as covariates contain indicators of the father’s social class, as well as father’s and mother’s income.

regression estimates for comparison.<sup>27</sup> Remark that the ATT estimate depends on the estimated  $\rho$ , hence on the set of instruments used, see Table 3.

When family, school and local controls are included (specification 3) the mean effect drops to 0.35 points when the maths and reading test scores (administered at age 7) are used as instruments, and the effect is insignificant from zero. The significance of the ATT estimate drops further when draw-a-man and copying designs are used instead, and the point estimate becomes zero when parental education is used as instrument. Similar results are obtained in the specifications that include additional LEA controls (column 4) or squares and interactions of covariates (column 5).

Hence the mean effect of selective schooling is estimated to be insignificant from zero when accounting for observables and unobservables. This contrasts with the methods that account for observables only (rows 1 to 4 in Table 4) and shows that the mean differences observed between comprehensive and selective schools are almost entirely driven by differences in composition, which are only partially corrected when accounting for differences in observables. Lastly, controlling for lagged test scores (“value added”) yields an overestimate of the effect of selective schooling, consistently with the discussion in Section 4.2, although the magnitude of the overestimation is moderate.

## 7.2 Distributional effects

We now turn to distributional effects. Panel b1) in Figure 1 shows the density of the score in mathematics in the selective system (solid line) and the comprehensive one (dashed), directly estimated on the raw data.<sup>28</sup> The density in the selective system is clearly bimodal, while the one in the comprehensive system presents a single mode. Moreover, as shown by panel a1), which plots the c.d.f.’s in the two systems, the distribution of outcomes in the comprehensive system is stochastically dominated by the one in the selective system. This means that children in the selective system do better at every quantile of the distribution.

The graphs on the second column of Figure 1 show the results when accounting for selection on observables and unobservables. The solid lines still represent the c.d.f. (top) and the density (bottom) of the realized outcome in the selective system, while the dashed lines now show the distribution of potential outcomes of the children attending

---

<sup>27</sup>Regression estimates are obtained by replacing, in (27), the conditional expectation by a linear projection, unconditional expectations by sample means, and  $\rho$  by  $\hat{\rho}$ .

<sup>28</sup>Densities and c.d.f.’s of realized outcomes are estimated using a Gaussian kernel, with Silverman’s rule of thumb as bandwidth choice.

a selective school, had they instead attended a comprehensive school. To estimate the latter, we use the nonparametric deconvolution approach that we introduced in Section 5. In Appendix B, we show how we choose the trimming parameter  $T_N$  in (29). The c.d.f. is then estimated by numerical integration of the density. In the graphs, we use the covariates specification which includes family, school and local characteristics, and we use draw-a-man and copying designs as instruments to estimate the ratio of returns to the endowment  $\rho$ .

Panel a2) in Figure 1 shows that the difference between the c.d.f.'s of potential outcomes in the two systems, for children who actually attended a selective school, is much reduced compared to the difference between the c.d.f.'s of realized outcomes. The reduction operates at every quantile, and visually suggests that the large differences observed in the raw data are largely due to composition effects. Once differences in composition are corrected for, the effect is zero below the percentile 60, and looks quantitatively small above that percentile. In addition, panel b2) shows that the estimated distribution of counterfactual outcomes is unimodal, unlike the distribution of realized outcomes in the selective system.<sup>29</sup>

To assess the order of magnitude of the differences between the two education systems at various points of the distribution, we plot in Figure 2 the quantile treatment effects  $\Delta(\tau)$  against the values of  $\tau \in [0, 1]$ . We present the results for two covariates specifications: family characteristics (column 1), and family, school and local characteristics (column 2). We also report the estimates obtained using either of the three choices of instruments to estimate  $\rho$ .

In all specifications, the quantile treatment effects are close to zero below the median. However, the various choices of instruments yield different effects above the median. In our preferred covariates specification (column 2), the effect is positive and significant from zero between the percentiles 70 and 90 when using lagged scores as instruments (maths and reading, or draw-a-man and copying designs). Yet, the effects remain small, and reach a peak at the percentile 80, where the quantile treatment effect is equal to 1 point in mathematics when using draw-a-man and copying designs as instruments. This is 15% of a standard deviation, and only one fourth of the difference in quantiles calculated on the raw data.

---

<sup>29</sup>To assess the bimodality or unimodality of a distribution, the choice of trimming for computing our nonparametric density estimator is crucial. In Appendix B we investigate the robustness of some of our results to variation in the choice of the trimming parameter.

This suggests that there may be a positive but small effect of attending a selective school above the median outcome. However, this effect is not robust to other specifications. Indeed, panels c1) and c2) in Figure 2 show that, when using father’s and mother’s education as instruments, we obtain zero effect above the median also. Hence, according to these results, the effects of selective schooling on maths outcomes are at best small and mostly insignificant, and we cannot reject that the effects are zero throughout the distribution.

**Grammar and secondary modern schools.** To interpret these distributional results, it is interesting to relate the estimated quantile treatment effects to the two types of selective secondary schools: grammar and secondary modern. In the rest of this subsection we focus on the estimation of the following mean difference:

$$\Delta^G = \mathbb{E}(Y_{i2}|G_i = 1) - \mathbb{E}(Y_{i2}^0|G_i = 1),$$

where  $G_i$  is the indicator of attending a secondary school of the grammar (i.e., academically demanding) type.  $\Delta^G$  measures the gain of attending a grammar school rather than a comprehensive school, for the children who actually attended a grammar school. We similarly define  $\Delta^S$ , the gain of attending a secondary modern school instead of a comprehensive one, for the children who actually attended a secondary modern school.

We estimate  $\Delta^G$  as

$$\hat{\Delta}^G = \frac{1}{N_G} \sum_{G_i=1} Y_{i2} - \frac{1}{N_G} \sum_{G_i=1} \hat{F}_{Y_{i2}^0|D_i=1}^{-1} \left[ \hat{F}_{Y_{i2}|D_i=1}(Y_{i2}) \right],$$

where  $\hat{F}_{Y_{i2}^0|D_i=1}$  is the estimated c.d.f. of counterfactual outcomes  $Y_{i2}^0$  given  $D_i = 1$ ,  $\hat{F}_{Y_{i2}|D_i=1}$  is the estimated c.d.f. of realized outcomes in the selective system, and  $N_G$  denotes the number of children attending a grammar school.<sup>30</sup> We proceed similarly to estimate  $\Delta^S$ .

For  $\hat{\Delta}^G$  to provide a consistent estimate of  $\Delta^G$ , a condition of *rank invariance* is needed. Rank invariance requires that the relative position of a child attending a selective school, among all children attending selective schools, would be the same if all children attending selective schools attended comprehensive schools instead. This is a strong assumption, which we do not need to estimate the distribution of  $Y_{i2}^0$  given  $D_i = 1$ ,

---

<sup>30</sup>In practice, we estimate the mean on the range 2%-98% of  $Y_{i2}$  given  $D_i = 1$  (i.e., between 2 and 28 points in mathematics), because the density of potential outcomes is not as well estimated in the tails.

but which is needed in order to identify the distributions of potential outcomes at the individual level, hence within grammar and secondary modern schools.<sup>31</sup>

Table 5 shows the estimates  $\widehat{\Delta}^G$  and  $\widehat{\Delta}^S$ , for various choices of instruments and covariates specifications. In our preferred specification (column 3 in the table) we find that the effect of attending a secondary modern school rather than a comprehensive school is insignificant from zero. This is consistent with Figure 2, which shows insignificant quantile treatment effects below the median outcome, where the outcomes of secondary modern students are concentrated.

As before, the various choices of instruments yield different results above the median outcome, hence different estimates of the effect of attending a grammar school. The effect is significant when using the maths and reading scores (age 7) as instruments to estimate  $\rho$ , and amounts to 0.8 points (11% of a standard deviation) in our preferred specification of covariates. However, the estimate drops to 0.6 points when using draw-a-man and copying design test scores as instruments, and becomes insignificant from zero. Lastly, the effect is negative, but insignificant, when using parental education as instrument. These differences can be compared to the raw differentials in test scores: children at grammar schools perform on average 8.6 points better than children in comprehensives, while children at secondary schools perform on average 1.5 points worse (compare Tables 1 and 2). We thus find that the raw differences are almost entirely due to the very large discrepancies in initial endowments between children in the three types of schools.

These results are related to Clark (2007), who focuses on the East Ridings district in the same period as us, and compares grammar and secondary schools directly. Controlling for the assignment test score (which is not available in the NCDS data) and using regression discontinuity and IV strategies, he finds small and mostly insignificant differences in test scores. Our results—obtained under the rank invariance assumption—are thus consistent with the evidence that he provides.<sup>32</sup>

---

<sup>31</sup>Rank invariance could fail to hold if some children had a comparative advantage in the selective system, for example because they benefited more than others from the positive externalities exerted by good students. If rank invariance does not hold, the second term in  $\widehat{\Delta}^G$  is a weighted average of outcomes in the selective school, that is not necessarily a consistent estimate of the counterfactual mean  $\mathbb{E}(Y_{i2}^0 | G_i = 1)$ .

<sup>32</sup>Note that, although Clark (2007) finds small effects on test scores, he finds larger differences in longer-run outcomes between children attending grammar and secondary modern schools.

### 7.3 Robustness checks

The above results suggest that the observed mean and distributional differences in test scores between selective and comprehensive schools are almost entirely due to differences in composition. Here we check the validity of this conclusion by performing various exercises.

**Multidimensional endowment.** First, we estimate the model allowing for two or three unobserved endowments. In the first case we use the test scores in mathematics and reading administered at age 11 as pre-secondary schooling outcomes (i.e., as variables  $Y_{i1}$  in the first part of the paper), while in the second case we add the verbal test score administered at 11. In both cases we use the maths, reading, draw-a-man and copying designs tests administered at age 7, as well as father's and mother's education, as instruments in order to estimate the parameter  $\rho$  (now a two or three-dimensional vector).

In our preferred specification of covariates (including family, school and local controls) we estimate the mean effect to be 0.07 points in mathematics, with a standard error of 0.32, when allowing for two unobserved endowments. Allowing for a third endowment, we find a negative mean effect of -0.14 with a standard error of 0.36.

Figure 3 show the quantile treatment effects. Allowing for two unobserved endowments yields a positive and significant, but quantitatively small, effect around the percentile 80, and slightly negative effects around the percentile 40. When allowing for a third endowment, the point estimates remain very similar, but the precision of the estimates worsens a lot. Overall, this evidence suggests that our results are reasonably robust to allowing for several unobserved endowments.

**Transformations of test scores.** Next, we estimate the effect of selective education on several transformations of test score variables. As explained in Section 4.3, by estimating the entire distribution of transformed outcomes we can also recover the distribution of outcomes in the original transformation, i.e., in terms of raw test scores. Then, from the estimated density we can recover the mean effect. Likewise, we can also estimate the mean effect of attending a grammar (resp. a secondary modern) school rather than a comprehensive school, as explained in the previous subsection.

Table 6 shows the results, for four transformations of the score in mathematics administered at age 16 (denoted as  $Y$ ): the logarithm  $\log(Y + 1)$ , a normal transform of the

percentiles  $\Phi^{-1}(F_{Y_{i2}}^{-1}(Y))$ ,  $\Phi$  being the standard normal c.d.f., and two Box-Cox transformations  $\frac{(Y+1)^{1/2}-1}{1/2}$  and  $\frac{(Y+1)^{3/2}-1}{3/2}$ . The same transformation is used for the maths test scores administered at age 16 ( $Y_{i2}$ ) and age 11 ( $Y_{i1}$ ).<sup>33</sup>

Most effects reported in Table 6 are similar to those estimated using the raw test scores. In particular, the average effect of attending a selective school is insignificant from zero in all specifications. The effects of attending a grammar school are of similar magnitudes as estimated in Table 5. However, they are now insignificant in every specification. The same holds for the effect of attending a secondary school.<sup>34</sup> Obtaining rather similar point estimates using transformed test scores suggests that our results are not driven by the assumption that the original test scores are linear in the endowment.

**Comprehensive schools.** Another concern is that the estimated effect of selective schooling includes schools that became comprehensive after 1969, as well as schools that were comprehensive before 1968, see Section 6. Hence, the data include pupils who started in a selective school, and ended their secondary education in a comprehensive school. The NCDS data provides the date when a school became comprehensive. We checked that, for 1288 out of the 3865 children whose school was comprehensive in 1974, the school was already comprehensive in 1969. Those children performed slightly worse in the age 16 test in mathematics compared to the children whose school became comprehensive after 1969 (11.5 points versus 12.1). Yet, they also performed worse at the maths exam administered at age 11 (14.5 points versus 16.3). Interestingly, when dropping the schools that became comprehensive after 1969 from the sample, our estimate of the ATT becomes close to zero irrespective of the instruments used.<sup>35</sup> Thus, it seems that children who started at a selective school which became comprehensive later on had better results at the age 16 test, only because they were already performing better at age 11.

**Comprehensive and selective LEAs.** We also tried to estimate the model using only LEAs which were purely selective or purely comprehensive in the period that we study.

---

<sup>33</sup>We used  $\log(Y+1)$  instead of  $\log(Y)$  to avoid the zeros in  $Y$ . Also, we used a normal transform of the percentiles instead of the percentiles themselves, as percentiles are distributed as a uniform, the density of which is difficult to estimate using kernel deconvolution methods. Lastly, note that applying the nonlinear IV strategy mentioned in Section 4.3, using all age 7 test scores as instruments, suggests that the log transformation is optimal.

<sup>34</sup>Indeed, there are no marked differences between the quantile treatment effects estimated using the raw test scores (see Figure 2) and those estimated using transformations of test scores. Results available upon request.

<sup>35</sup>Results are available upon request.

Manning and Pischke (2006) argue that this provides a better comparison, as it avoids the possibility of endogenous school choice within a LEA. We focused on LEAs where either more than 90% of secondary schools were comprehensive (in 1972), or less than 10% of schools were comprehensive. Applying our method that accounts for observables and unobservables, we also found very little difference between the two education systems, although the smaller sample size (1884 observations) yielded more imprecise estimates.

**Pre-test.** Lastly, we measure the effects of attending a selective school on test scores administered *before* starting secondary education. As argued by Manning and Pischke (2006), finding a strong effect would suggest that our approach does not fully correct for the correlation between unobservables and the education system attended. The outcomes are the scores in mathematics, reading and verbal tests administered at age 11, and the covariates specification includes family characteristics. The results are presented in Table 7.

The two first rows in the table show the matching estimates of the average treatment effect, controlling for family characteristics only (row 1) and adding the age 7 scores in mathematics and reading as controls (row 2). This latter specification corresponds to the “value-added” methodology used in many papers, see Section 4.2. We find strong effects on the three scores, which drop by roughly 40% when including lagged test scores as controls, but remain significant.

The third row in Table 7 shows the ATT estimates, using our method to control for observables and the unobserved endowment. The first principal component of the maths and reading test scores administered at age 7 is used as pre-treatment outcome (variable  $Y_{i1}$ ), and father’s and mother’s education are used as instruments to estimate  $\rho$ . Indeed, age 7 test scores such as draw-a-man are now contemporaneous to  $Y_{i1}$ , so they are likely to be invalid instruments for  $Y_{i1}$  in (14). Using our methodology accounting for observables and unobservables, we find much smaller point estimates, insignificant from zero for the three outcomes. This suggests that, while the “value-added” strategy fails to fully control for differences in composition between the two education systems, our approach does allow to control for these differences.

## 8 Conclusion

We have proposed a method to compare the educational performance of pupils attending selective and non selective (or comprehensive) schools. The model specifies test scores as the output of an additive production function, of which the child's initial endowment is an essential, though unobserved, input. We have extended the logic of Difference-in-Differences (DID), which allows to estimate average effects, to estimate the entire counterfactual distribution of the potential outcomes of pupils of selective schools, had they instead attended a comprehensive school. The estimators of the density and quantile treatment effects that we propose are related to nonparametric deconvolution. We have also shown how to allow the returns to the endowment before and after secondary schooling to be different.

Applying the methodology to NCDS data, we have found that the average effect of attending a selective school is at best very small, and mostly insignificant. Moreover, although we find some evidence of positive effects at the top of the distribution of outcomes, these effects are quantitatively small and are not robust across specifications. Hence, our analysis suggests that the raw differences in performance between selective and non selective schools are almost entirely due to differences in pupils' composition.

The methodology adopted in this paper could be useful in other contexts. Our results apply to models of the form:

$$Y_{it}(D_{it}) = f_t(X_{it}, D_{it}) + \eta_i + v_{it}(D_{it}),$$

where  $Y_{it}(0)$  and  $Y_{it}(1)$  denote the potential outcomes of a binary treatment  $D_{it} \in \{0, 1\}$ . Our approach allows to estimate the entire distribution of  $Y_{it}(0)$  and  $Y_{it}(1)$  (on the treated), if errors  $(v_{it}(0), v_{it}(1))$  are independent of  $D_{it}$  and  $\eta_i$ , while the endowment  $\eta_i$  can be correlated to  $X_{it}$  and  $D_{it}$  in an unrestricted way.

Our approach and the one proposed by Athey and Imbens (2006) are non nested. On the one hand, our method is more restrictive as additivity is assumed, and the estimator is not invariant to monotone transformations of the outcomes. On the other hand, we leave the distribution of  $v_{it}(D_{it})$  unrestricted and allow for general distributional effects. Hence, the present paper provides a useful alternative to Athey and Imbens' work in contexts where additivity is motivated by economic assumptions on the technology generating the outcomes.

## References

- [1] ABADIE, A. (2005): “Semiparametric Difference-in-Differences Estimators,” *Review of Economic Studies*, 72, 1-19.
- [2] ABREVAYA, J. (2002): “Computing Marginal Effects in the Box-Cox Model,” *Econometric Reviews*, 21(3), 383-393.
- [3] AMEMIYA, T. (1985): *Advanced Econometrics*. Cambridge, MA: Harvard University Press.
- [4] ATHEY, S., and G. IMBENS (2006): “Identification and Inference in Nonlinear Difference-in-Differences Models,” *Econometrica*, 74(2), 431-497.
- [5] CARNEIRO, P., K. HANSEN, and J. HECKMAN (2003): “Estimating Distributions of Treatment Effects with an Application to the Returns to Schooling and Measurement of the Effects of Uncertainty on College Choice,” *International Economic Review*, 44(2), 361-422.
- [6] CARROLL, R. J., and P. HALL (1988): “Optimal rates of Convergence for Deconvoluting a Density,” *Journal of the American Statistical Association*, 83, 1184-1186.
- [7] CHITTY, C. (2002): “The Role and Status of LEAs: post-war pride and *fin de siècle* uncertainty,” *Oxford Education Review*, 28, 2-3, 247-260.
- [8] CLARK, D. (2007): “Selective Schools and Academic Achievement,” *IZA* working paper n. 3192.
- [9] CONNOLLY, S., J. MICKLEWRIGHT, and S. NICKEL (1992): “The Occupational Success of Young Men Who Left School at Sixteen,” *Oxford Economic Papers*, 44, 460-479.
- [10] CROOK, D. (2002): “Local Authorities and Comprehensivisation in England and Wales: 1944-1974,” *Oxford Education Review*, 28, 2-3, 247-260.
- [11] CUNHA, F., and J. HECKMAN (2006): “Formulating, Identifying and Estimating the Technology of Cognitive and Noncognitive Skill Formation ,” *Mimeo*.
- [12] CUNHA, F., J. HECKMAN, and S. SCHENNACH (2006): “Estimating the Technology of Cognitive and Noncognitive Skill Formation,” *Mimeo*.

- [13] DEARDEN, L., J. FERRI, and C. MEGHIR (2002): “The Effect of School Quality on Educational Attainment and Wages,” *Review of Economics and Statistics*, 84, 1-20.
- [14] DIGGLE, P. J., and P. HALL (1993): “A Fourier Approach to Nonparametric Deconvolution of a Density Estimate,” *Journal of the Royal Statistical Society Series B*, 55, 523-531.
- [15] FAN, J.Q. (1991): “On the Optimal Rates of Convergence for Nonparametric Deconvolution Problems,” *Annals of statistics*, 19, 1257-1272.
- [16] FIRPO, S. (2007): “Efficient Semiparametric Estimation of Quantile Treatment Effects,” *Econometrica*, 75, 259-276.
- [17] GALINDO-RUEDA, F., and A. VIGNOLES (2005): “The Heterogeneous Effect of Selection In Secondary Schools: Understanding the Changing Role of Ability,” *Working Paper, Center for the Economics of Education*.
- [18] HANSEN, K., J. HECKMAN, and K. MULLEN (2004): “The Effect of Schooling and Ability on Achievement Test Scores,” *Journal of Econometrics*, 121, 39-98.
- [19] HANUSHEK, E., and L. WOESSMAN (2006): “Does Educational Tracking Affect Performance and Inequality ? Differences-in-Differences Evidence Across Countries,” *Economic Journal*, 116, C36-C76.
- [20] HAYDN, T. (2004): “The Strange Death of the Comprehensive School in England and Wales, 1965-2002”, *Research Papers in Education*, 19, 415-432.
- [21] HECKMAN, J.J. (1990), “Varieties of Selection Bias,” *American Economic Review*, 80, 313-318.
- [22] HECKMAN, J.J., J.N. SMITH, and N. CLEMENTS (1997), “Making the Most Out of Program Evaluations and Social Experiments: Accounting for Heterogeneity in Program Impacts,” *Review of Economic Studies*, 64, 487-536.
- [23] HECKMAN, J., J. STIXRUD and S. URZUA (2006): “The Effect of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior,” *Journal of Labor Economics*, 24(39), 411-482.

- [24] HIRANO, K., G. IMBENS and G. RIDDER (2003): “Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score,” *Econometrica*, 71(4), 1161-1189.
- [25] HONORE, B., and E. TAMER (2006): “Bounds on Parameters in Dynamic discrete-Choice Models,” *Econometrica*, 74(3), 611-629.
- [26] HOLTZ-EAKIN, D., W. NEWEY and H. ROSEN (1988): “Estimating Vector Autoregressions with Panel Data,” *Econometrica*, 56(6), 1371-1395.
- [27] KERCKHOFF, A.C. (1986): “Effects of Ability Grouping in British Secondary Schools,” *American Sociological Review*, 51, 842-858.
- [28] KERCKHOFF, A.C., K. FOGELMAN, D. CROOK, and D. REEDER (1996): *Going Comprehensive in England and Wales: A study of Uneven Change*. London and Portland: The Woburn Press.
- [29] LINDGREN, B.W. (1993): *Statistical Theory*, Chapman & Hall, New York.
- [30] MANNING, A., and J.S. PISCHKE (2006): “Comprehensive versus Selective Schooling in England and Wales: What do we Know?,” *NBER Working Paper*, n.12176.
- [31] MAURIN, E., and S. McNALLY (2006): “Selective Schooling,” *Mimeo*.
- [32] MEGHIR, C., and M. PALME (2005): “Educational Reform, Ability and Family Background,” *American Economic Review*, 95(1), 414-424.
- [33] ROSENBAUM, P. and D. RUBIN (1983): “The Central Role of the Propensity Score in Observational Studies for Causal Effects,” *Biometrika*, 70, 41-55.
- [34] RUBIN, D. (1974): “Estimating Causal Effects of Treatments in Randomized and Non-Randomized Studies,” *Journal of Educational Psychology*, 66, 688-701.
- [35] SCHAFFER, J.L. (1997): *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall.
- [36] THUYSBAERT, B. (2007): “Distributional Comparisons in Difference-in-Differences Models,” *mimeo*.

- [37] TODD, P.E. and K.I. WOLPIN (2003): "On the Specification and Estimation of the Production Function of Cognitive Achievement ," *Economic Journal*,113, F3-F33.
- [38] TODD, P.E. and K.I. WOLPIN (2004): "The Production of Cognitive Achievement in Children: Home, School and Racial Test Score Gaps," *University of Pennsylvania Working Paper*.

# APPENDIX

## A Proofs

**Proof of Theorem 1** Assumption 2 implies that, for  $t \in \mathbb{R}$ :

$$\begin{aligned}\Psi_{Y_{i2}^0|D_i=1}(t) &= \exp(j\alpha_{20}t) \Psi_{\eta_i|D_i=1}(t) \Psi_{v_{i2}^0|D_i=1}(t) \\ \Psi_{Y_{i2}^0|D_i=0}(t) &= \exp(j\alpha_{20}t) \Psi_{\eta_i|D_i=0}(t) \Psi_{v_{i2}^0|D_i=0}(t).\end{aligned}$$

Assumption 1 then implies that

$$\begin{aligned}\Psi_{Y_{i2}^0|D_i=1}(t) &= \exp(j\alpha_{20}t) \Psi_{\eta_i|D_i=1}(t) \Psi_{v_{i2}^0}(t) \\ \Psi_{Y_{i2}^0|D_i=0}(t) &= \exp(j\alpha_{20}t) \Psi_{\eta_i|D_i=0}(t) \Psi_{v_{i2}^0}(t).\end{aligned}$$

So:

$$\Psi_{Y_{i2}^0|D_i=1}(t) = \frac{\Psi_{\eta_i|D_i=1}(t)}{\Psi_{\eta_i|D_i=0}(t)} \Psi_{Y_{i2}^0|D_i=0}(t),$$

which is well-defined by Assumption 3.

Lastly, using Assumptions 2, 1 and 3 in turn we get similarly:

$$\frac{\Psi_{Y_{i1}|D_i=1}(t)}{\Psi_{Y_{i1}|D_i=0}(t)} = \frac{\Psi_{\eta_i|D_i=1}(t)}{\Psi_{\eta_i|D_i=0}(t)}.$$

This ends the proof.

**Proof of Theorem 2.** The proof of (10) is very similar to that of Theorem 1. Indeed:

$$\begin{aligned}\Psi_{Y_{i2}^0|D_i=1}(t) &= \mathbb{E} \left[ \Psi_{Y_{i2}^0|D_i=1, X_i}(t|X_i) | D_i = 1 \right] \\ &= \int \Psi_{Y_{i2}^0|D_i=1, X_i}(t|X_i) dP(X_i | D_i = 1) \\ &= \mathbb{E} \left[ \frac{p_D(X_i)}{p_D} \Psi_{Y_{i2}^0|D_i=1, X_i}(t|X_i) \right] \\ &= \mathbb{E} \left[ \frac{p_D(X_i)}{p_D} \frac{\Psi_{Y_{i1}|D_i=1, X_i}(t|X_i)}{\Psi_{Y_{i1}|D_i=0, X_i}(t|X_i)} \Psi_{Y_{i2}^0|D_i=0, X_i}(t|X_i) \right] \\ &= \frac{1}{p_D} \mathbb{E} \left[ \omega(t|X_i) (1 - p_D(X_i)) \Psi_{Y_{i2}^0|D_i=0, X_i}(t|X_i) \right] \\ &= \frac{1}{p_D} \mathbb{E} \left[ \omega(t|X_i) (1 - D_i) \exp(jtY_{i2}) \right],\end{aligned}$$

where going from the second to the third line requires use of Bayes' rule, and the last equality comes from applying the law of iterated expectations.

## B Choice of the trimming parameter

The parameter  $T_N$  ensures that the integrals converge in (25). In order to choose  $T_N$  in practice, we use a rule of thumb along the lines of Diggle and Hall (1993). In a simple deconvolution problem, Diggle and Hall show that an optimal  $T_N$  must satisfy  $\Psi_X(T_N) = N^{-1/2}$ , where  $\Psi_X$  is the characteristic function of the random variable  $X$ , the distribution of which is unknown

and is to be estimated. We checked that:  $\log|\widehat{\Psi}_{Y_{i2}^0|D_{i=1}}(t)|$  is almost linear in  $t^2$  over a wide range. See Figure A1 for an illustration. We extrapolate the estimated characteristic function outside of this range and solve for

$$\left|\widehat{\Psi}_{Y_{i2}^0|D_{i=1}}(T_N)\right| = N^{-1/2}$$

on the basis of this extrapolation. Doing so provided reasonable guesses for the bandwidth. Figure A2 shows the effect of varying the trimming parameter  $T_N$  from .42, the choice suggested by our informal method, to 1. We see that the oscillations increase when  $T_N$  increases. However, the estimates of the quantile treatment effects  $\Delta(\tau)$  are not much affected.

It is beyond the scope of this paper to derive the asymptotic properties of the estimators. Nevertheless, we conjecture that the bootstrap is valid, and we use nonparametric bootstrap in order to compute asymptotic standard errors (100 replications).

## C Variables

**Covariates specifications used.** Throughout the paper, we use the following lists of covariates:

*Family characteristics* include the gender of the child, father’s and mother’s education, number of older siblings, all the father’s social class measures available in the 1958, 1965 and 1969 survey rounds, measures of father’s and mother’s income (only banded/categorized income measures are available in the NCDS), and whether the mother is working in 1965.

*School characteristics* include the number of children in the child’s class (1965), the number of schools attended (1969), the pupil teacher ratio (1969), whether the child is in junior school in 1969, the age of the main school buildings (1969), and dummies for an ability streamed class in 1969.

*Local characteristics* include various proportions at the ward level, obtained from the census data: proportion of unemployed and sick, proportion of mining workers, of working mothers, proportions of skilled, semi-skilled, managerial and unskilled workers, proportion of households with indoor WC, and proportion of immigrants, as well as regional dummies.

*Additional LEA controls* include LEA-level variables obtained by matching the NCDS with the dataset on the “Effect of Local Education Authority Resources and Policies on Educational Attainment, 1972-1974”, number SN199, available from the UK data archive.<sup>36</sup> These are the proportion of years (between 1957 and 1970) during which the LEA was controlled by the Labour party, the population size and density of each LEA, and an industrialisation index, as well as the proportions of 13 year old girls and boys in comprehensive schools in 1967 and in 1972. The data are matched on the NCDS identifiers from the school’s questionnaire. We do not use the information on medical exams or other tests as these could have been conducted in different locations in some instances. We construct variables on the percentages of comprehensive schools (in 1967 and 1972) in the LEA where the child was living in 1969.

**Draw-a-man and copying designs.** The “draw-a-man” test is a non-verbal measure of ability. The test score ranges from 0 to 59 and is conceived after Goodenough. It usually consists of requiring children to draw a man, a woman and themselves on separate pieces of paper. Scoring is based on details of the drawings and additional points are given for the level of detail, presence and proportion of the parts drawn. In the NCDS test, children were only required to make a single drawing of a man.

---

<sup>36</sup>[www.data-archive.ac.uk](http://www.data-archive.ac.uk)

The “copying designs” test requires the child to make two attempts each at copying given basic shapes (circle, cross, triangle, etc.) onto a piece of paper. Also the child is asked to copy a simple sentence on paper. The test is designed to test the motoric coordination and precision.

**Missing values.** Most of the covariates in the NCDS have some missing values. To address this problem, we adopted the following simple approach:<sup>37</sup> for each regressor we replaced the missing values by 0, and we created a dummy variable that takes the value one if that regressor is missing. The percentages of missing values range between 2% and 20%. Note that we only proceeded in this way for the variables that we use as controls, given that they are not the main focus of the analysis. Hence missing values of test scores are not treated in this way. Also, when using father’s and mother’s education as instruments, we used only observations with non-missing values for those two variables.

## D Evidence on mobility

In this section of the appendix, we provide some evidence on between-LEA mobility, and mobility to private schools. For this purpose, we use the original NCDS data, merged with the dataset on the “Effect of Local Education Authority Resources and Policies on Educational Attainment, 1972-1974”, number SN199 (see previous section).

**School changes.** Table A1 shows some descriptive statistics on comprehensive and selective schools, using the full NCDS sample that includes children who changed school between 1969 and 1974. We find very similar results to the tests compared to the subsample of children who stayed during five years in the same school, which we use in the analysis.

**Between-LEA mobility.** Here we document mobility between LEAs. According to the NCDS dataset, between-LEA mobility is less frequent when the child is in secondary school (between 11 and 16 years old, 11.5% of observations) than in primary school (between 7 and 11 years old, 21.7%). Then we relate the probability of changing LEA to family and LEA characteristics. For this we construct for each child the shares (measured in 1967 and 1972) of comprehensive schools in the LEA where he lived in 1965, irrespective of whether the child changed LEA between 1965 and 1969. We obtain two variables: the share of comprehensive schools in 1967 in the LEA of origin, and the variation in that share between 1967 and 1972. We proceed similarly for children aged 11, and obtain the share of comprehensive schools in 1967 in the LEA of origin (now corresponding to the LEA where the child lived in 1969), and the variation in that share between 1967 and 1972.

Table A2 shows the estimates of a probit regression of the indicator of moving between LEAs between 1965 and 1969 on the share of comprehensives and the variation in that share, as well as the estimates of a probit regression of the indicator of moving between LEA between 1969 and 1974. We find that parental education is positively associated with geographic mobility. Family and LEA characteristics together explain little of the variance between age 7 and 11 (pseudo  $R^2$  is 2%), but their explanatory power increases between age 11 and 16 (pseudo  $R^2$  is 8%). Interestingly, LEAs of origin with a larger share of comprehensive schools are associated with *less* mobility. The same holds for LEAs with an increasing share of comprehensive schools (between 1967 and 1972). In both cases the effects are small and close to insignificant. These results do not provide evidence that parents reacted to the comprehensivisation reform by moving to another LEA. Remark that this is consistent with comprehensive and selective schools

---

<sup>37</sup>A more rigorous approach would be to use multiple imputation (see, e.g., Schafer, 1997).

having roughly similar effects on pupils performance, which is the conclusion of the present study.

**Mobility to private schools.** Next, we consider private schools.<sup>38</sup> The percentage of private secondary schools is 6.7% in the NCDS. Descriptive statistics are given in Table A3. Pupils in private schools perform much better in all tests than pupils in comprehensive or selective schools. Comparison with Table 2 shows that children in private schools do slightly worse than children in grammar schools, although they have on average a much better parental background.

Next, we regress by probit the indicator variable that a child attended a private school (measured in 1974) on family characteristics and the share of comprehensives in the LEA where the child was living in 1969. We also include an interaction term with the first principal component of age 11 test scores (column 2). We find that family characteristics explain a large share of the variance (pseudo  $R^2$  of 21%). However, the share of comprehensive schools has an insignificant effect on the probability of attending a private school. This suggests that focusing on non-private schools will not bias the comparison of comprehensive and selective schools, as long as we control for parental characteristics.

---

<sup>38</sup>We define private schools as “non-LEA” schools, for which we have information on the LEA where the child lived in 1974.

Table 1: Descriptive statistics, selective and comprehensive schools

Variable	Comprehensive			Selective		
	Mean	Std.Dev.	N	Mean	Std.Dev.	N
Maths score (age 16)	11.9	6.3	3600	13.8	7.1	2872
Maths score (age 11)	15.7	9.6	3434	18.7	10.4	2636
Reading score (age 11)	15.7	5.9	3434	17.1	6.0	2636
Verbal score (age 11)	21.6	9.0	3436	24.2	9.1	2636
Maths score (age 7)	5.1	2.4	3431	5.4	2.4	2701
Reading score (age 7)	23.1	6.9	3437	24.4	6.4	2717
Draw-a-man score (age 7)	24.0	6.8	3373	24.7	6.9	2648
Copying designs score (age 7)	7.1	1.9	3426	7.3	1.9	2710
Father's education	3.8	1.6	2997	4.1	1.8	2392
Mother's education	3.9	1.3	3037	4.0	1.4	2427

Table 2: Descriptive statistics, grammar and secondary modern schools

Variable	Grammar			Secondary Modern		
	Mean	Std.Dev.	N	Mean	Std.Dev.	N
Maths score (age 16)	20.5	5.2	985	10.4	5.3	1887
Maths score (age 11)	28.5	6.4	913	13.4	8.0	1723
Reading score (age 11)	22.1	4.4	913	14.5	5.0	1723
Verbal score (age 11)	32.0	4.8	913	20.2	8.1	1723
Maths score (age 7)	6.9	2.1	936	4.7	2.3	1765
Reading score (age 7)	28.6	2.2	939	22.1	6.7	1778
Draw-a-man score (age 7)	27.3	6.6	915	23.3	6.7	1733
Copying designs score (age 7)	8.0	1.7	934	6.9	1.9	1776
Father's education	4.7	2.1	824	3.7	1.6	1568
Mother's education	4.5	1.7	834	3.7	1.1	1593

Table 3:  $\rho$  estimates, using various sets of instruments

Instruments		(1)	(2)	(3)	(4)
Mathematics and reading scores (age 7)	$\rho$ estimate	0.520 (0.015)	0.522 (0.015)	0.525 (0.015)	0.530 (0.015)
	Shea's partial $R^2$	0.361	0.357	0.357	0.357
	F-statistic	761	759	754	749
	Sargan p-value	0.404	0.438	0.280	0.578
Draw-a-man and copying designs scores (age 7)	$\rho$ estimate	0.561 (0.023)	0.563 (0.023)	0.562 (0.023)	0.555 (0.025)
	Shea's partial $R^2$	0.133	0.129	0.129	0.135
	F-statistic	219	217	207	207
	Sargan p-value	0.135	0.132	0.151	0.119
Father's and mother's education	$\rho$ estimate	0.662 (0.058)	0.683 (0.063)	0.682 (0.069)	0.702 (0.080)
	Shea's partial $R^2$	0.024	0.021	0.018	0.018
	F-statistic	37	34	31	21
	Sargan p-value	0.383	0.403	0.353	0.337

Note: 2-Stage Least Squares regression of the score in mathematics at 16 on the score in mathematics at 11 on the subsample of children attending a comprehensive school, using various sets of instruments. Specification (1) includes family characteristics, (2) and (3) include in addition school, and school and local controls, respectively, see Appendix C. Specification (4) is (3) plus squares and interactions. Standard errors clustered at the LEA level in parentheses.

Table 4: ATT estimates of attending a selective school on the score in mathematics at age 16

		Selection on observables				
		(1)	(2)	(3)	(4)	(5)
Regression		1.539 (0.278)	1.536 (0.272)	1.487 (0.296)	1.272 (0.315)	1.497 (0.295)
Matching		1.555 (0.267)	1.514 (0.292)	1.449 (0.408)	0.789 (0.701)	1.261 (0.466)
Regression (including age 11 maths score)		0.505 (0.167)	0.533 (0.168)	0.434 (0.175)	0.378 (0.192)	0.439 (0.178)
Matching (including age 11 maths score)		0.628 (0.207)	0.610 (0.225)	0.429 (0.388)	-0.043 (0.743)	0.473 (0.295)
		Selection on observables and unobservables				
		(1)	(2)	(3)	(4)	(5)
	Instruments					
Regression	Maths and reading scores (age 7)	0.410 (0.176)	0.446 (0.172)	0.349 (0.199)	0.256 (0.228)	0.362 (0.237)
	Copying designs and draw-a-man scores (7)	0.350 (0.189)	0.392 (0.191)	0.301 (0.199)	0.228 (0.249)	0.350 (0.212)
	Father's and mother's education	-0.006 (0.255)	0.010 (0.235)	-0.085 (0.317)	-0.153 (0.299)	-0.098 (0.303)
Matching	Maths and reading scores (age 7)	0.404 (0.175)	0.433 (0.173)	0.349 (0.227)	0.084 (0.304)	0.252 (0.268)
	Copying designs and draw-a-man scores (7)	0.345 (0.187)	0.385 (0.194)	0.289 (0.222)	0.023 (0.311)	0.241 (0.230)
	Father's and mother's education	0.001 (0.255)	0.021 (0.231)	-0.080 (0.303)	-0.274 (0.302)	-0.135 (0.262)

Note: Regression and matching estimates of the effect of attending a selective school, average effect on the treated. Regression and matching estimates are described in the text. Specification (1) includes family characteristics, (2), (3) and (4) include in addition school, school and local, and school, local and additional LEA controls, respectively, see Appendix C. Specification (5) is (3) plus squares and interactions. Bootstrapped standard errors clustered at the LEA level in parentheses (100 replications).

Table 5: ATT estimates of attending a grammar (resp. a secondary modern) school on the score in mathematics at age 16

Grammar versus comprehensive					
Instruments	(1)	(2)	(3)	(4)	(5)
Mathematics and reading (age 7)	0.940 (0.247)	0.920 (0.283)	0.796 (0.413)	0.280 (0.455)	0.660 (0.461)
Draw-a-man and copying designs (age 7)	0.869 (0.266)	0.817 (0.269)	0.586 (0.366)	0.177 (0.560)	0.696 (0.631)
Father's and mother's education	-0.089 (0.561)	-0.304 (0.662)	-0.342 (0.609)	-0.688 (0.722)	-0.458 (0.769)
Secondary modern versus comprehensive					
Instruments	(1)	(2)	(3)	(4)	(5)
Mathematics and reading	0.311 (0.177)	0.358 (0.168)	0.339 (0.336)	-0.095 (0.422)	0.227 (0.404)
Draw-a-man and copying designs	0.294 (0.178)	0.344 (0.198)	0.240 (0.275)	-0.051 (0.503)	0.217 (0.600)
Father's and mother's education	-0.066 (0.253)	-0.047 (0.272)	-0.101 (0.251)	-0.289 (0.456)	0.155 (0.388)

Note: Estimates of the effect of attending a grammar school versus attending a comprehensive school ( $\Delta^G$ ), and of the effect of attending a secondary modern school versus attending a comprehensive school ( $\Delta^S$ ), average effects on the treated. The estimates are computed from the estimated counterfactual density of outcomes, see Section 7.2. Covariates specifications are as in Table 4. Bootstrapped standard errors clustered at the LEA level in parentheses (100 replications).

Table 6: Average treatment effects based on various transformations of the test scores in mathematics (age 16)

Instruments	Transformations			
	$\log(Y + 1)$	$\Phi^{-1}(F_Y(Y))$	$\frac{(Y+1)^{1/2}-1}{1/2}$	$\frac{(Y+1)^{3/2}-1}{3/2}$
Mean effect of attending a selective school ( $\Delta$ )				
Math and reading	0.367 (0.272)	0.654 (0.548)	0.433 (0.287)	-0.204 (0.256)
Draw-a-man and copying designs	0.262 (0.221)	0.520 (0.362)	0.349 (0.247)	-0.184 (0.261)
Father's and mother's education	-0.171 (0.337)	0.026 (0.413)	-0.144 (0.407)	-0.523 (0.431)
Mean effect of attending a grammar school ( $\Delta^G$ )				
Math and reading	0.508 (0.375)	0.659 (0.742)	0.640 (0.428)	0.221 (0.328)
Draw-a-man and copying designs	0.287 (0.337)	0.495 (0.566)	0.530 (0.381)	0.209 (0.324)
Father's and mother's education	-0.328 (0.553)	-0.157 (0.632)	-0.257 (0.709)	-0.572 (0.655)
Mean effect of attending a secondary modern school ( $\Delta^S$ )				
Math and reading	0.305 (0.272)	0.444 (0.427)	0.282 (0.235)	-0.147 (0.197)
Draw-a-man and copying designs	0.263 (0.198)	0.339 (0.294)	0.197 (0.210)	-0.129 (0.202)
Father's and mother's education	-0.071 (0.270)	-0.067 (0.305)	-0.136 (0.321)	-0.241 (0.285)

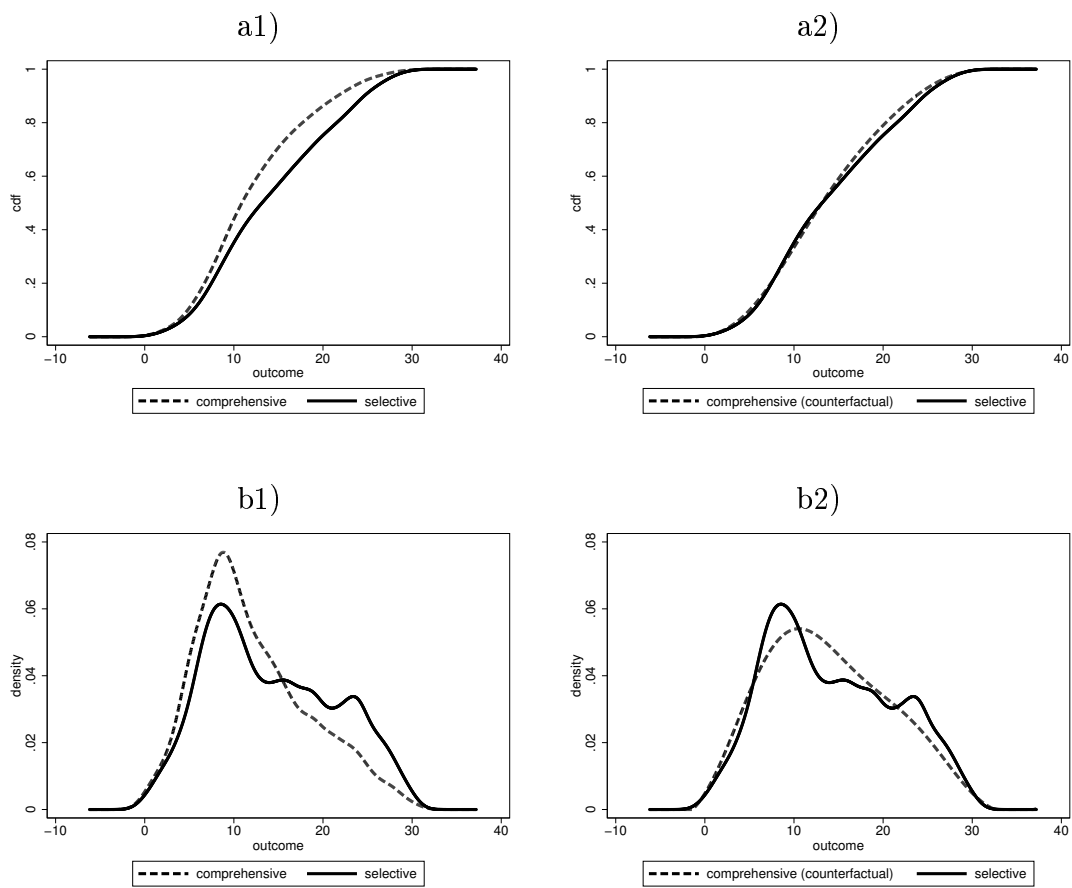
Note: The counterfactual density of outcomes in the comprehensive system is recovered from the estimated counterfactual density of transformed outcomes. From that density, mean effects are then computed. Covariates specification includes family, school and local characteristics. Bootstrapped standard errors clustered at the LEA level in parentheses (100 replications).

Table 7: ATT on test scores administered at age 11

	Mathematics	Reading	Verbal
Controlling for observables	2.222 (0.454)	1.011 (0.275)	2.063 (0.467)
Controlling for observables and age 7 test scores	1.464 (0.330)	0.579 (0.238)	1.326 (0.413)
Controlling for observables and unobservables	0.321 (0.501)	0.105 (0.356)	0.415 (0.625)

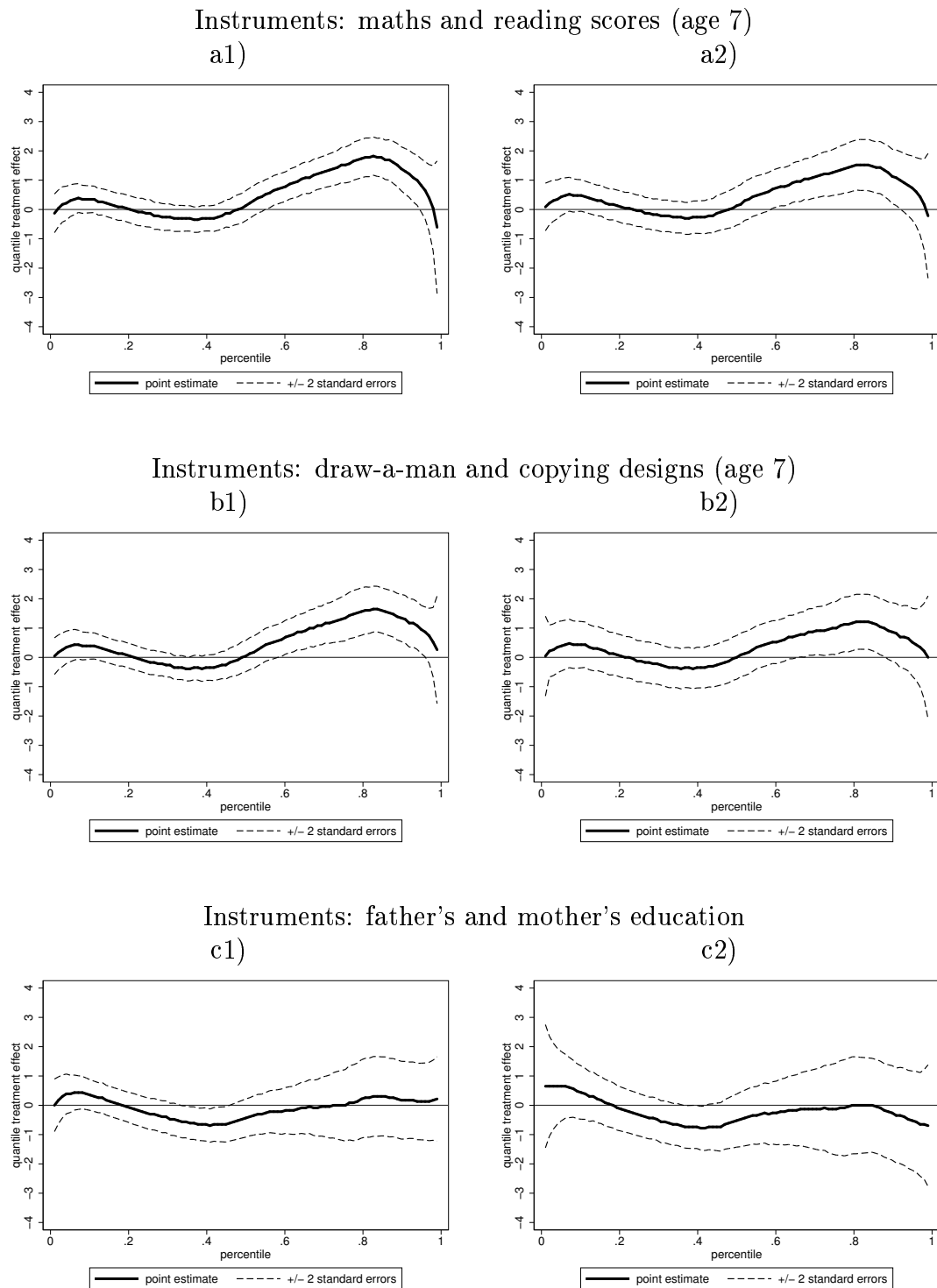
Note: Average treatment effect on the treated of attending a selective rather than a comprehensive school on age 11 test scores. Covariates specification includes family characteristics, as well as age 7 reading and mathematics test scores in the second row. Father's and mother's education are used as instruments in the third row. Bootstrapped standard errors clustered at the LEA level in parentheses (100 replications).

Figure 1: Distributional effects of attending a selective school on the maths score at 16



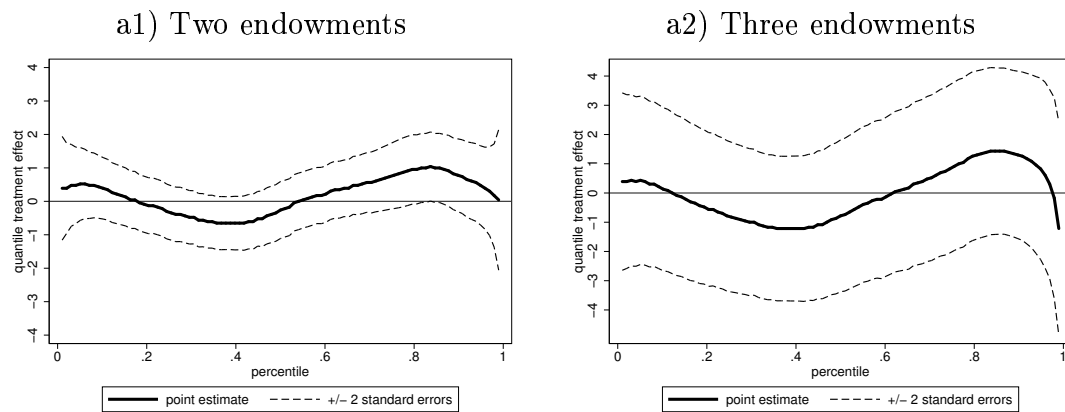
Note: Estimates of the cumulative distribution function (top) and density (bottom) of realized and counterfactual outcomes. C.d.f.'s/densities of realized outcomes are estimated using a Gaussian kernel. C.d.f.'s/densities of potential outcomes are estimated by integrating characteristic function estimates, see Section 5. Draw-a-man and copying designs are used as instruments to estimate  $\rho$ . Covariates specification includes family, school and local characteristics.

Figure 2: Quantile treatment effects of attending a selective school on the maths score administered at age 16



Note: Quantile treatment effects  $\Delta(\tau)$  on the y-axis,  $\tau \in [0, 1]$  on the x-axis. First column: covariates specification includes family characteristics. Second column: covariates specification includes family, school and local characteristics. Various sets of instruments are used to estimate  $\rho$ . Solid lines show point estimates, dashed lines show confidence bands of  $\pm 2$  bootstrapped standard errors, clustered at the LEA level (100 replications).

Figure 3: Quantile treatment effects allowing for two and three endowments

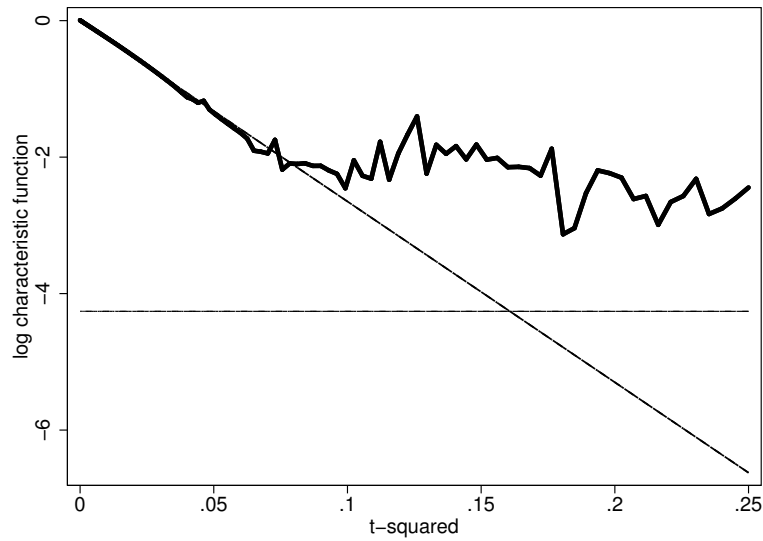


Note: Quantile treatment effects  $\Delta(\tau)$  on the y-axis,  $\tau \in [0, 1]$  on the x-axis, allowing for 2 or 3 unobserved endowments. The instruments used to estimate  $\rho$  are: maths, reading, draw-a-man and copying designs (age 7) and father's and mother's education. Covariates specification includes family, school and local characteristics. Solid lines show point estimates, dashed lines show confidence bands of  $\pm 2$  bootstrapped standard errors, clustered at the LEA level (100 replications).

Table A1: Descriptive statistics, selective and comprehensive schools in the full sample

Variable	Comprehensive			Selective		
	Mean	Std.Dev.	N	Mean	Std.Dev.	N
Maths score (age 16)	11.8	6.3	5406	13.7	7.1	3666
Maths score (age 11)	15.6	9.7	5114	18.5	10.5	3307
Reading score (age 11)	15.5	5.9	5114	17.0	6.1	3308
Verbal score (age 11)	21.4	8.9	5116	24.1	9.1	3308
Maths score (age 7)	5.0	2.4	5159	5.4	2.4	3412
Reading score (age 7)	23.0	6.9	5168	24.3	6.4	3428
Draw-a-man score (age 7)	24.0	6.8	5074	24.6	7.0	3347
Copying designs score (age 7)	7.1	1.9	5153	7.3	1.9	3420
Father's education	3.9	1.7	4451	4.1	1.9	3064
Mother's education	3.9	1.3	4526	4.0	1.4	3109

Figure A1: Selection of the trimming parameter



Note: Graph of  $\log \left| \widehat{\Psi}_{Y_2^0 | D_i=1}(t) \right|$  as a function of  $t^2$ .  $T_N$  can be found at the intersection of the two straight lines ( $T_N^2 \approx .16$ ). Draw-a-man and copying designs are used as instruments to estimate  $\rho$ . Covariates specification includes family, school and local characteristics.

Table A2: Between-LEA mobility, 1965-1969 and 1969-1974

	1965-1969		1969-1974	
	(1)	(2)	(3)	(4)
% comprehensives (1967) in LEA of origin	-0.0007 (0.001)	-0.0009 (0.002)	-0.0012 (0.0008)	-0.0017 (0.001)
Variation in % comprehensives (1967-1972) in LEA of origin		-0.003 (0.002)		0.0012 (0.0014)
Mathematics score age 7 (age 11 in columns 3 and 4)	-0.010 (0.009)	-0.010 (0.009)	0.008 (0.004)	0.0082 (0.0036)
Reading score age 7 (age 11 in columns 3 and 4)	0.009 (0.003)	0.009 (0.004)	0.008 (0.005)	0.0073 (0.0048)
Father's education	0.029 (0.019)	0.03 (0.019)	0.026 (0.011)	0.025 (0.011)
Mother's education	0.032 (0.014)	0.032 (0.014)	0.032 (0.015)	0.034 (0.015)
Pseudo R-squared	0.02	0.02	0.08	0.08

Note: Probit regression of the variable indicating if the child changed LEA between 1965 and 1969 (columns 1 and 2) or between 1969 and 1974 (columns 3 and 4). Controls for family characteristics included. Standard errors clustered at the LEA level in parentheses.

Table A3: Descriptive statistics: private schools

Variable	Mean	Std.Dev.	N
Maths score (age 16)	18.7	6.7	681
Maths score (age 11)	25.8	9.5	545
Reading score (age 11)	21.6	5.5	545
Verbal score (age 11)	28.8	7.7	545
Maths score (age 7)	6.7	2.3	563
Reading score (age 7)	27.2	4.6	564
Draw-a-man score (age 7)	26.3	7.3	555
Copying designs score (age 7)	7.6	1.8	560
Father's education	6.0	2.5	449
Mother's education	5.6	2.0	453

Table A4: Attending a private secondary school

	(1)	(2)
% Comprehensive in the LEA	-0.0008 (0.0012)	-0.0015 (0.0014)
1st principal component of age 11 test scores (pcscores)	0.229 (0.029)	0.207 (0.042)
pcscores $\times$ % comp.		0.0008 (0.0011)
Father's education	0.048 (0.018)	0.048 (0.018)
Mother's education	0.100 (0.019)	0.100 (0.019)
Pseudo R-squared	0.21	0.22

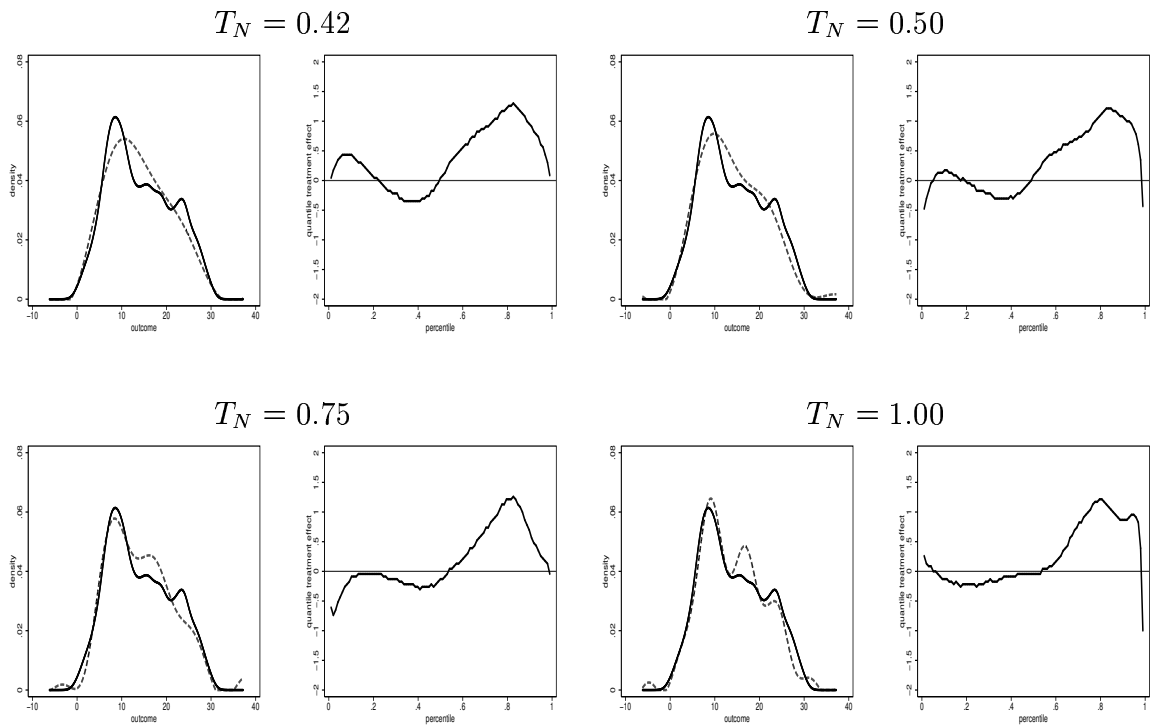
Note: Probit regression of the variable indicating if the child attended a private school in 1974. Control for family characteristics included. Standard errors clustered at the LEA level in parentheses.

Table A5: Sample correlations

Maths score (age 16)	1.00									
Maths score (age 11)	0.76	1.00								
Reading score (age 11)	0.60	0.68	1.00							
Verbal score (age 11)	0.60	0.72	0.73	1.00						
Maths score (age 7)	0.44	0.51	0.43	0.46	1.00					
Reading score (age 7)	0.43	0.51	0.57	0.63	0.48	1.00				
Draw-a-man score (age 7)	0.27	0.30	0.30	0.34	0.29	0.31	1.00			
Copying designs score (age 7)	0.31	0.33	0.27	0.31	0.27	0.27	0.34	1.00		
Father's education	0.28	0.27	0.25	0.23	0.12	0.15	0.13	0.11	1.00	
Mother's education	0.28	0.26	0.25	0.23	0.14	0.16	0.11	0.09	0.45	1.00

Note: Sample is restricted to children who have non-missing observations for all variables ( $N = 4081$ ).

Figure A2: Choice of the trimming parameter and distribution estimates



Note: Estimates of the density of realized (solid line) and counterfactual outcomes (dashed), and quantile treatment effects estimates. The density of realized outcomes is estimated using a Gaussian kernel. Densities of potential outcomes are estimated by integrating characteristic function estimates, for various choices of the trimming parameter  $T_N$ . Draw-a-man and copying designs are used as instruments to estimate  $\rho$ . Covariates specification includes family, school and local characteristics.