

Recovering Distributions in Difference-in-Differences Models: A Comparison of Selective and Comprehensive Schooling¹

Stéphane Bonhomme²
CEMFI, Madrid

Ulrich Sauder³
University of Warwick

Revised version: August 2009

¹We 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.

²**Corresponding author:** CEMFI, Casado del Alisal, 5, 28014 Madrid, Spain; bonhomme@cemfi.es

³Department of Economics, Coventry CV47AL, England; usauder@gmx.net

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. Test scores are modelled as the output of an additive production function. An important input is the child's unobserved initial endowment, which may be correlated with the education system attended. In this model, we generalize the Difference-in-Differences approach and identify the entire counterfactual distribution of potential outcomes. Our results suggest that the better performance of selective schools relative to non selective ones is essentially due to differences in pupils' composition.

JEL codes: C33, I21.

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

1 Introduction

Should pupils be segregated according to their performance, or should all pupils be grouped together? Given the diversity of education systems around the world and the absence of a clear theoretical answer to this question, empirical measures of the effect of selective education on children’s outcomes are needed.¹

For this purpose, the UK experience in the 1970s is of special value as two systems of secondary education coexisted for some time. 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.

Several researchers have tried to compare the comprehensive and selective systems in the UK, controlling for differences in parental background or characteristics of the primary school.² Most 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 likely to be correlated with unobservables that affect later outcomes.

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 the UK 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.

Attending a selective or comprehensive secondary school is assumed not to depend on the shocks to future educational attainment (between ages 11 and 16). However, the distributions of initial endowments of children in the two education systems are allowed to be different. For this reason, the hypothesis of selection on observables (e.g., Rosenbaum and Rubin, 1983) does not hold in the model: if children about to attend a selective school are better endowed before starting, differences at age 16 between 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, AI hereafter) also provide identification results for distributions of potential outcomes in a DID framework. However, the assumptions needed in their approach 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, Athey and Imbens's approach fails at recovering the true distribution in general. As an example, we show that when distributions are normal AI's approach will be inconsistent unless the distribution of period 2 outcomes is a mean shift of the distribution of out-

comes in period 1, or the endowment is independent of the treatment.³ In contrast, our approach allows the distributions of pre- and post-treatment outcomes to have different shapes.

We consider three extensions of this 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.⁴ 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, which we identify using lagged outcomes as instruments. Lastly, we discuss how to estimate mean and distributional effects if some transformations of test scores (instead of the test scores themselves) are linear in the 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. When continuous covariates are present, we propose to use the inverse probability weighting method of Hirano, Imbens and Ridder (2003) and Abadie (2005) for estimating average effects. To estimate the entire distribution of outcomes nonparametrically, we use a standard kernel 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.

We then apply the methodology to the NCDS sample. Descriptive statistics show that children perform better in selective than in comprehensive schools. Our results show that these differences are essentially due to differences in pupils’ composition. When accounting for differences in observables and unobservables, the mean effect becomes small and insignificant. Quantile effects are also small, at most 15% of a standard deviation at percentile 80 where the effect is maximum. Various robustness checks confirm the results.

The outline of the paper is as follows. In Section 2 we present the model and study the identification of distributions. Section 3 discusses estimation and inference. We then present the empirical application in Section 4, and conclude in Section 5.

2 Identification analysis

In this section we present the model that we will use to quantify the effects of selective education, and study the identification of the distribution of potential outcomes.

2.1 A model of test scores

There are two periods: before secondary education (period 1, age 11 in our empirical application), and after (period 2, age 16). Let Y_{i1} be a test score in period 1, and Y_{i2} be a test score measured in period 2, where i indexes individual units. Let also $D_i = 1$ (respectively $D_i = 0$) denote 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. 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 a potential—outcome.

We suppose the following model test scores:

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

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 16.⁵ So, our comparison of 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 η_i , which contains the child’s cognitive ability. We allow η_i to be correlated with X_i and D_i , as parental inputs

and school choice may be based on the child’s ability. 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.

To conduct the analysis, we restrict the production function to be 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 β_1 and β_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.⁶ At the end of this section we discuss how 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.⁷

Allowing for the presence of η_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, η_i , which is not observed by the econometrician, acts as a “confounder” and complicates the estimation. As the methods previously proposed in the treatment effects literature do not apply, we need to develop new strategies for identifying and estimating these effects. We now discuss the identification of the distribution of Y_{i2}^0 given $D_i = 1$ (*on the treated*), starting with a special case.

2.2 Identification in a special case

In the particular case without covariates and with equal returns to the endowment, 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 α_1 and α_2^0 are scalar parameters.

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.⁸

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 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 η_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 η_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 η_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 average treatment effect on the treated (ATT) is then given by:

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

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 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 η_i given D_i .

Assumption 2 requires the shocks to the outcomes to be independent of the individual-specific endowment η_i . This for example rules out the possibility that the shocks have individual-specific variances. Note that η_i is allowed to be correlated with the treatment

D_i , so η_i is analogous to a “fixed effect” in a panel data model as it is independent of time-varying innovations but may be correlated with D_i . Moreover, the shocks to test scores v_{i1} and v_{i2}^0 are allowed to be correlated in an unrestricted way.

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 .⁹ A one-to-one mapping with the density function¹⁰ of W , say f_W , is given by the inverse Fourier transformation:

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

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

It is very common in the nonparametric deconvolution literature to assume that some characteristic functions have no real zeros (see Schennach, 2004, for references). Many usual parametric distributions (normal, Gamma...) satisfy this condition. Examples of distributions with characteristic functions having real zeros are the uniform and the symmetrically truncated normal.¹¹

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:

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

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 η_i . Moreover, correcting for differences in η_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.¹²

Having obtained the identification of the characteristic function of potential outcomes, the identification of their density immediately follows by taking the inverse Fourier transformation.

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

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

Corollary 1 thus shows that the entire distribution of potential outcomes is identified. So, in addition to the ATT, the quantile treatment effects defined as follows are also identified:

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

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.¹³

Relationship with Athey and Imbens (2006). Our approach is related to Athey and Imbens' (2006, AI hereafter) "Changes-in-Changes" (CIC) model. The approach in AI relies on the following identity for the c.d.f. of potential outcomes:

$$F_{Y_{i2}^0|D_i=1}(y) = F_{Y_{i1}|D_i=1} \left(F_{Y_{i1}|D_i=0}^{-1} \left(F_{Y_{i2}|D_i=0}(y) \right) \right). \quad (10)$$

An important remark for our purpose is that (10) is generally *not* satisfied in model (3) under our assumptions.

To illustrate this point, let us consider the special case where the unobservables follow normal distributions. Then (10) is not satisfied, unless the distributions of v_{i1} and v_{i2}^0

are identical, or η_i is independent of D_i . To see why, let η_i be normally distributed given $D_i = k$, with mean and variance μ_k and σ_k^2 ($k = 0, 1$), and let v_{i1} and v_{i2}^0 be normal with zero mean and variance ν_1^2 and ν_2^2 , respectively. Then, simple algebra¹⁴ shows that the right-hand side of (10) is the c.d.f. of a normal distribution with mean $\alpha_2^0 + \xi\mu_1 + (1-\xi)\mu_0$ and variance $\xi^2(\sigma_1^2 + \nu_1^2)$, where $\xi = \sqrt{\frac{\sigma_0^2 + \nu_2^2}{\sigma_0^2 + \nu_1^2}}$. Those coincide with the mean and variance of $Y_{i2}^0 | D_i = 1$ ($\alpha_2^0 + \mu_1$ and $\sigma_1^2 + \nu_2^2$, respectively) only if $\nu_1^2 = \nu_2^2$ or $(\mu_0, \sigma_0^2) = (\mu_1, \sigma_1^2)$, i.e. if the distributions of v_{i1} and v_{i2}^0 are identical, or if η_i is independent of D_i .

The reason why AI may fail to recover the distribution of potential outcomes in model (3) is that their approach requires the unobservables to have the same distribution in the two periods, for treated and controls (see AI's Assumption 3.3, p.439). In the context of our application this restriction is unappealing,¹⁵ 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, allowing for example comprehensive schools to have an equalizing effect on children's outcomes.

It is important to emphasize that the generality of our approach is obtained at the cost of imposing additivity on the outcome variables. Although not unnatural in production function applications, additivity is often difficult to justify from economic theory. In addition, 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.¹⁶

2.3 Identification in the general case

We now 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.

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. We now 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. Assumptions 1, 2 and 3 are assumed valid conditional on X_i . This covers cases where potential outcomes are additive in X_i ,

as in equations (2). 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 time-varying shocks and the treatment, yet they leave the correlation between η_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 η_i into account when deciding what primary school to choose (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 $D_i = 1$. As in the case without covariates, knowledge of the characteristic function implies knowledge of the density, using the inverse Fourier transformation.

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

$$\Psi_{Y_{i2}^0|D_i=1, X_i}(t|x) = \frac{\Psi_{Y_{i1}|D_i=1, X_i}(t|x)}{\Psi_{Y_{i1}|D_i=0, X_i}(t|x)} \Psi_{Y_{i2}|D_i=0, X_i}(t|x), \quad (11)$$

and

$$\Psi_{Y_{i2}^0|D_i=1}(t) = \frac{1}{p_D} \mathbb{E}[\omega(t|X_i) (1 - D_i) \exp(jtY_{i2})], \quad (12)$$

where we have denoted as

$$\begin{aligned} \omega(t|X_i) &\equiv \frac{p_D(X_i)}{(1 - p_D(X_i))} \frac{\Psi_{Y_{i1}|D_i=1, X_i}(t|X_i)}{\Psi_{Y_{i1}|D_i=0, X_i}(t|X_i)} \\ &= \frac{\mathbb{E}[D_i \exp(jtY_{i1}) | X_i]}{\mathbb{E}[(1 - D_i) \exp(jtY_{i1}) | X_i]}. \end{aligned} \quad (13)$$

It is interesting to remark that in model (3) potential outcomes are *not* independent of the treatment given observables, i.e. selection on observables (Rosenbaum and Rubin, 1983) 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 η_i . As the distribution of η_i may differ between treated and controls, estimators based on the selection-on-observables assumptions will be biased in general. The empirical part will illustrate that taking into account differences in unobservables may be very important in schooling applications.

Allowing for different returns to unobservables. The benchmark model (3) imposes that the coefficients of η_i in the equations of pre- and post-treatment outcomes are the same. In some instances, one may want to allow for different coefficients, for example to allow ability to be differently rewarded at age 11 and 16, and to 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:

$$\begin{aligned} Y_{i2}^0 &= \alpha_2^0 + \beta_2^0 \eta_i + v_{i2}^0, \\ Y_{i1} &= \alpha_1 + \beta_1 \eta_i + v_{i1}, \end{aligned} \quad (14)$$

where α_1 , β_1 , α_2^0 and β_2^0 are scalar parameters. Note that (14) implies that

$$Y_{i2}^0 = \alpha_2^0 - \rho \alpha_1 + \rho Y_{i1} + v_{i2}^0 - \rho v_{i1}, \quad (15)$$

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

Y_{i1} is endogenous in equation (15), 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:*

$$\begin{cases} v_{i1} \text{ and } v_{i2}^0 \text{ are uncorrelated with } \tilde{Y}_{i0} \text{ given } D_i = 0, \\ Y_{i1} \text{ and } \tilde{Y}_{i0} \text{ are correlated given } D_i = 0. \end{cases} \quad (16)$$

Importantly, \tilde{Y}_{i0} is not assumed independent of η_i or of potential outcomes.¹⁷ Thus \tilde{Y}_{i0} is an ‘‘instrument’’ in the sense of the literature on linear panel data models, being orthogonal to the time-varying innovations though not to the fixed effect. In the application to schooling, we will use lagged test scores to instrument Y_{i1} in (15), in a similar way as in previous work on panel data (e.g., Holtz-Eakin *et al.*, 1988).

Under Assumption 5, \tilde{Y}_{i0} is a valid instrument for Y_{i1} in (15) when conditioning on $D_i = 0$. So ρ is identified as:

$$\rho = \frac{\text{Cov}(\tilde{Y}_{i0}, Y_{i2}^0 | D_i = 0)}{\text{Cov}(\tilde{Y}_{i0}, Y_{i1} | D_i = 0)}. \quad (17)$$

Provided that ρ is identified, the above analysis may be replicated, yielding:

$$f_{Y_{i2}^0 | D_i=1}(y) = \frac{1}{2\pi} \int \exp(-jty) \left[\frac{\Psi_{Y_{i1} | D_i=1}(\rho t)}{\Psi_{Y_{i1} | D_i=0}(\rho t)} \Psi_{Y_{i2}^0 | D_i=0}(t) \right] dt. \quad (18)$$

Hence the identification of the entire distribution of potential outcomes.

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

Lastly, remark that our approach is *not* equivalent to conditioning on the first period's outcome, which is sometimes referred to as the "value-added" methodology. Indeed, the "value-added" estimand of the mean is:

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

This is different from the mean potential outcome in model (14):

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

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:

$$\begin{aligned} h(Y_{i2}^0; \lambda_2^0) &= \alpha_2^0 + \beta_2^0 \eta_i + v_{i2}^0, \\ h(Y_{i1}; \lambda_1) &= \alpha_1 + \beta_1 \eta_i + v_{i1}, \end{aligned} \tag{19}$$

where α_1 , β_1 , α_2^0 and β_2^0 are scalar parameters, λ_1 and λ_2^0 are vectors of parameters, and h is a known function that we suppose increasing in y .

We assume that λ_1 and λ_2^0 are known.¹⁸ 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.

3 Estimation and inference

In this section we discuss estimation of mean and distributional effects. A STATA program is available online.¹⁹

3.1 No covariates

We start with the simple model (3) without covariates, where the coefficients of η_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:

$$\hat{f}_{Y_{i2}^0 | D_i=1}(y) = \frac{1}{2\pi} \int_{-T_N}^{T_N} \exp(-jty) \left[\frac{\hat{\Psi}_{Y_{i1} | D_i=1}(t)}{\hat{\Psi}_{Y_{i1} | D_i=0}(t)} \hat{\Psi}_{Y_{i2} | D_i=0}(t) \right] dt. \quad (20)$$

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 (20) is finite. To guarantee the consistency of the estimator, T_N needs to tend to infinity with the sample size N .

To interpret the estimator given by (20), it is useful to reformulate the problem of estimating the density of potential outcomes as a nonparametric deconvolution problem. To do so, denote as $Y_{i2}(0)$ a random variable distributed as $(Y_{i2} | D_i = 0)$, and define similarly $Y_{i2}^0(1)$, $Y_{i1}(0)$, and $Y_{i1}(1)$. Let also $Y_1 \oplus Y_2$ denote the independent sum of Y_1

and Y_2 , distributed as the sum of independent draws from Y_1 and Y_2 . Remark that model (3) implies the following equality in distributions (see Theorem 1):

$$Y_{i2}(0) \oplus Y_{i1}(1) \stackrel{d}{=} Y_{i2}^0(1) \oplus Y_{i1}(0). \quad (21)$$

The distributional identity (21) may be interpreted as a nonparametric deconvolution problem, as its left-hand side is the sum of observed random variables, while its right-hand side is the independent sum of the latent factor $Y_{i2}^0(1)$, and the “error” $Y_{i1}(0)$. The distribution of the latter is not known *a priori*, yet a random sample is available from it. Hence, the problem of recovering the distribution of $Y_{i2}^0(1)$ in (21) is a deconvolution problem with *unknown* error distribution.

This observation allows us to derive the asymptotic properties of the density estimator given by (20). Convergence rates of deconvolution estimators in the case where the error distribution is known have been studied in several important papers (e.g., Carroll and Hall, 1988, and Fan, 1991a). In a recent paper, Johannes (2009) extends those results to the case where the error distribution is unknown and must be estimated. The kernel deconvolution estimator he considers coincides with (20), taking the product $\widehat{\Psi}_{Y_{i2}|D_i=0}(t)\widehat{\Psi}_{Y_{i1}|D_i=1}(t)$ as an estimator of the characteristic function of the independent sum $Y_{i2}(0) \oplus Y_{i1}(1)$.²⁰ As a consequence, his results may be applied to our setting.

Johannes (2009) finds that, as in the case of a known error distribution, the density estimator is consistent and its rate of convergence depends on the tails of the characteristic functions that appear in (20). Moreover, those rates may be very slow, even logarithmic when the tails of the characteristic function of errors are very thin and thus a low T_N is needed to ensure consistency. Slow rates of convergence are the price to pay in our nonparametric approach, where the dependence of D_i on η_i is left unrestricted.

The link between our approach and nonparametric deconvolution allows us to discuss inference. In the case where the error distribution is known, the kernel deconvolution estimator is asymptotically normal under suitable conditions (e.g., Fan, 1991b). Asymptotic normality has also been shown in cases where the error distribution is estimated (Horowitz and Markatou, 1996). In our empirical application, we will use the nonparametric bootstrap in order to compute pointwise confidence bands. Proving the consistency of the bootstrap is difficult in this context. In a recent paper, Bissantz *et al.* (2007) provide conditions under which the nonparametric bootstrap is consistent in the model with known error distribution. We conjecture that the bootstrap remains consistent when the error distribution needs to be estimated, although we are not aware of a

formal proof of this result.

Lastly, it is easy to extend the estimator in (20) to cases where the returns to η_i are different in the two periods. We propose to proceed in two steps. First, we estimate ρ 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:

$$\hat{f}_{Y_{i2}^0|D_i=1}(y) = \frac{1}{2\pi} \int_{-T_N}^{T_N} \exp(-jty) \left[\frac{\hat{\Psi}_{Y_{i1}|D_i=1}(\hat{\rho}t)}{\hat{\Psi}_{Y_{i1}|D_i=0}(\hat{\rho}t)} \hat{\Psi}_{Y_{i2}|D_i=0}(t) \right] dt. \quad (22)$$

Inference is performed by bootstrapping $\hat{\rho}$ and the density estimator simultaneously.

3.2 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 ρ , 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. Following Hirano *et al.* (2003) and Abadie (2005), the ATT can be shown to be equal to

$$\mathbb{E}(Y_{i2}|D_i = 1) - \mathbb{E}(Y_{i2}^0|D_i = 1) = \frac{1}{p_D} \mathbb{E} \left\{ \frac{p_D(X_i)}{1 - p_D(X_i)} (D_i - p_D(X_i)) (Y_{i2} - \rho Y_{i1}) \right\}. \quad (23)$$

We estimate the unconditional ATT as an empirical analog of (23). 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 ρ in (23) by $\hat{\rho}$.²¹ Note also that, for (23) to be well-defined, we need the propensity score to be strictly lower than 1. In the empirical application of Section 4, we select the observations for which the propensity score is between its 5th and 95th percentiles.

The counterfactual density of outcomes is estimated as:

$$\hat{f}_{Y_{i2}^0|D_i=1}(y) = \frac{1}{2\pi} \int_{-T_N}^{T_N} \exp(-jty) \frac{1}{\hat{p}_D} \left(\frac{1}{N} \sum_{i=1}^N \hat{\omega}(\hat{\rho}t|X_i) (1 - D_i) \exp(jtY_{i2}) \right) dt, \quad (24)$$

where $\hat{\omega}(t|X_i)$ is an estimate of $\omega(t|X_i)$ given by (13). To compute $\hat{\omega}(t|X_i)$ we replace the conditional expectations that appear at the numerator and denominator in the last line

of equation (13) by linear projections.²² We use a numerical approximation to compute the integral in (24).²³ 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. Then, once the density is estimated, the c.d.f. is directly obtained by numerical integration, and quantiles are computed by inversion of the estimated c.d.f. Lastly, pointwise confidence bands are computed using the nonparametric bootstrap.²⁴

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.

4 Comprehensive and selective schooling in the UK

In this section, we apply the approach to compare comprehensive and selective schooling in the UK in the 1970s, using the NCDS sample. More details on the data and additional results are available in the working paper version (Bonhomme and Sauder, 2009, BS hereafter).

4.1 Data description

The NCDS is an ongoing longitudinal survey of a British birth cohort born between March 3 and 9 of 1958. 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 (6.7% of the observations in the original NCDS data). In addition, we only keep children who attended the same school during their five years of secondary schooling. We checked, using the full NCDS data, that the way we restricted the sample is unlikely to bias the results (see BS for details).

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. We also use test score variables measured before starting secondary school: mathematics and reading at ages 7 and 11, verbal at age 11, and two additional tests administered at age 7: “draw-a-man” and “copying designs”.

The control variables that we use can be divided into three categories. Family char-

acteristics 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 (i.e., age 16) 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 local education authority (LEA) as an additional control, as well as additional LEA characteristics.²⁵ References about the data sources and the variables used are given in BS.

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

The selective system is by construction very heterogeneous. The second part of Table 1 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.

4.2 Estimation results

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 effect is to estimate the ratio of the returns to the unobserved

endowment η_i between age 11 and age 16, $\rho = \beta_2^0/\beta_1$. We estimate ρ 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 as instruments. 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 Subsection 4.1), which are less correlated with the scores in mathematics. 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.²⁶ Table 2 presents the results, for various specifications of covariates.

The ρ 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 ρ 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) Δ . The first two rows in Table 3 show the effects obtained when accounting for selection on observables only (computed using the inverse probability weighting method of Hirano *et al.*, 2003), while the next three rows show the effect ATT estimates, when accounting for differences in observables and unobservables between the two education systems (computed using the approach outlined above).

In our favorite covariates specification (3), which includes family, school and local controls, the estimates that do not correct for selection on unobservables (first row) are about 1.5 points in mathematics. When correcting also for selection on unobservables, the mean effect drops to 0.35 points when the maths and reading test scores administered at age 7 are used as instruments to estimate ρ , and the effect is insignificant from zero (third row).²⁷ 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 and shows that the raw mean differences between comprehensive and selective schools are almost entirely driven by differences in composition, which are only partially corrected for when accounting for differences in observables.

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.²⁸ 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 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 3. In Appendix B, we show how we choose the trimming parameter T_N in (24). 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 ρ .

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.

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 ρ .

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 significantly different 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.

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 this paragraph 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. Δ^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 Δ^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 Δ^G as

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

where $\widehat{F}_{Y_{i2}^0|D_i=1}$ is the estimated c.d.f. of counterfactual outcomes Y_{i2}^0 given $D_i = 1$, $\widehat{F}_{Y_{i2}|D_i=1}$ 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.²⁹ We proceed similarly to estimate Δ^S .

For $\widehat{\Delta}^G$ to provide a consistent estimate of Δ^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.³⁰ This is a strong assumption in the context of heterogeneous treatment effects. Importantly, we did not need rank invariance to estimate the distribution of Y_{i2}^0 given $D_i = 1$, although we use it here in order to identify the distributions of potential outcomes in grammar and secondary modern schools.³¹

Table 4 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 ρ , 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 (see Table 1). 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.³²

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. In the working paper version (Bonhomme and Sauder, 2009), we check the validity of this conclusion by performing various exercises.

First, we verify that allowing for a multidimensional unobserved endowment (two or three components) has little effect on the results. Next, we estimate the effect of selective education on several transformations of test score variables. As explained in Subsection 2.3, 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. Using a wide range of Box-Cox transformations, we obtained rather similar point estimates, suggesting that our results are not driven by the assumption that the original test scores are linear in the endowment.

As a last check, we measure the effects of attending a selective school on test scores administered *before* starting secondary education (in mathematics, reading, and verbal subjects at age 11). 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 results are shown in Table 5. The second row in the table shows that the “value-added” strategy that controls for the test score taken at age 7 fails to fully control for differences in composition between the two education systems, as it delivers a positive and significant effect of selective education on outcomes measured before entering secondary school. In contrast, the results in the third row, though not very precisely estimated, suggest that our approach does allow to control for those differences, as the estimated effect is insignificant from zero.

5 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) 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. This important conclusion is consistent with the findings of Manning and Pischke (2006), and casts doubt on previous measures of the effect of selective education on educational performance.

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 η_i , while the endowment η_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] Abadie, A., J. Angrist, and G. Imbens (2002), "Instrumental Variables Estimates of the Effect of Subsidized Training on the Quantiles of Trainee Earnings," *Econometrica*, 70, 91-117.

- [3] Abrevaya, J. (2002), "Computing Marginal Effects in the Box-Cox Model," *Econometric Reviews*, 21(3), 383-393.
- [4] Amemiya, T. (1985), *Advanced Econometrics*. Cambridge, MA: Harvard University Press.
- [5] Athey, S., and G. Imbens (2006), "Identification and Inference in Nonlinear Difference-in-Differences Models," *Econometrica*, 74(2), 431-497.
- [6] Bissantz, N., L. Dumbgen, H. Holzmann, and A. Munk (2007), "Nonparametric Confidence Bands in Deconvolution Density Estimation," *J. R. Stat. Soc. Ser. B*, 69, 483-506.
- [7] Bonhomme, S., and U. Sauder (2009), "Accounting for Unobservables in Comparing Selective and Comprehensive Schooling ," CEMFI working paper 0906.
- [8] 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.
- [9] Carrasco, M., and J. P. Florens (2009), "Spectral Methods for Deconvolving a Density," to appear in *Econometric Theory*.
- [10] Carroll, R. J., and P. Hall (1988), "Optimal rates of Convergence for Deconvoluting a Density," *Journal of the American Statistical Association*, 83, 1184-1186.
- [11] Clark, D. (2007), "Selective Schools and Academic Achievement," *IZA* working paper n. 3192.
- [12] Cunha, F., and J. Heckman (2008), "Formulating, Identifying and Estimating the Technology of Cognitive and Noncognitive Skill Formation ," *Journal of Human Resources*, 43(4), 738-782.
- [13] Cunha, F., J. Heckman, and S. Schennach (2006), "Estimating the Technology of Cognitive and Noncognitive Skill Formation," unpublished working paper.
- [14] 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.

- [15] 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.
- [16] Fan, J.Q. (1991a), "On the Optimal Rates of Convergence for Nonparametric Deconvolution Problems," *Annals of statistics*, 19, 1257-1272.
- [17] Fan, J.Q. (1991b), "Asymptotic Normality for Deconvolution Kernel Density Estimators," *Sankhya Ser. A*, 53, 97-110.
- [18] 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*.
- [19] Hall, P., and S.N. Lahiri (2008), "Estimation of Distributions, Moments and Quantiles in Deconvolution Problems," *Annals of Statistics*, 36, 2110-2134.
- [20] Hansen, K., J. Heckman, and K. Mullen (2004), "The Effect of Schooling and Ability on Achievement Test Scores," *Journal of Econometrics*, 121, 39-98.
- [21] Hanushek, E., and L. Woessman (2006), "Does Educational Tracking Affect Performance and Inequality ? Differences-in-Differences Evidence Across Countries," *Economic Journal*, 116, C36-C76.
- [22] Heckman, J.J. (1990), "Varieties of Selection Bias," *American Economic Review*, 80, 313-318.
- [23] 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.
- [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] Holtz-Eakin, D., W. Newey and H. Rosen (1988), "Estimating Vector Autoregressions with Panel Data," *Econometrica*, 56(6), 1371-1395.
- [26] Horowitz, J. L., and M. Markatou (1996), "Semiparametric Estimation of Regression Models for Panel Data," *Review of Economic Studies*, 63, 145-168.

- [27] Johannes, J. (2009), "Deconvolution with Unknown Error Distribution," *Annals of Statistics*, 37, 2301-2323.
- [28] Kerckhoff, A.C. (1986), "Effects of Ability Grouping in British Secondary Schools," *American Sociological Review*, 51, 842-858.
- [29] Manning, A., and J.S. Pischke (2006), "Comprehensive versus Selective Schooling in England and Wales: What do we Know?," *NBER Working Paper*, n.12176.
- [30] Maurin, E., and S. McNally (2006), "Selective Schooling," unpublished working paper.
- [31] Meghir, C., and M. Palme (2005), "Educational Reform, Ability and Family Background," *American Economic Review*, 95(1), 414-424.
- [32] Rosenbaum, P. and D. Rubin (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41-55.
- [33] Rubin, D. (1974), "Estimating Causal Effects of Treatments in Randomized and Non-Randomized Studies," *Journal of Educational Psychology*, 66, 688-701.
- [34] Schennach, S. (2004), "Estimation of Nonlinear Models with Measurement Error," *Econometrica*, 72, 33-75.
- [35] Thuysbaert, B. (2007), "Distributional Comparisons in Difference-in-Differences Models," unpublished working paper.
- [36] 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.
- [37] 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 By a standard property of characteristic functions, Assumption 2 implies that, for $t \in \mathbb{R}$:

$$\Psi_{Y_{i2}^0|D_i=1}(t) = \exp(j\alpha_2^0 t) \Psi_{\eta_i|D_i=1}(t) \Psi_{v_{i2}^0|D_i=1}(t).$$

So, using Assumption 1:

$$\Psi_{Y_{i2}^0|D_i=1}(t) = \exp(j\alpha_2^0 t) \Psi_{\eta_i|D_i=1}(t) \Psi_{v_{i2}^0}(t).$$

Likewise, we have:

$$\Psi_{Y_{i2}|D_i=0}(t) = \exp(j\alpha_2^0 t) \Psi_{\eta_i|D_i=0}(t) \Psi_{v_{i2}^0}(t).$$

Combining both identities yields:

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

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 (11) is very similar to that of Theorem 1. Indeed:

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

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 (20). 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 Ψ_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. 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. In our experiments we found that although the choice of T_N may affect the shape of the estimated density, the estimates of quantile treatment effects are not much affected in general. See Bonhomme and Sauder (2009) for more details.

Notes

¹Recent examples of studies aiming at measuring the effect of selective education on educational performance are Meghir and Palme (2005), Maurin and McNally (2006), and Hanushek and Woessman (2006).

²Examples are Kerckhoff (1986), Dearden *et al.* (2002), and Galindo-Rueda and Vignoles (2005).

³In a recent paper, Thuysbaert (2007) considers an additive version of the Athey and Imbens (2006) model where the distribution of the single unobservable does not change over time. This implies that, conditional on observables, the distribution of potential outcomes in period 2 is a mean shift of that in period 1. Under this condition, he constructs root- N consistent estimators of functionals of the distributions of potential outcomes.

⁴This 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).

⁵Allowing 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. The characteristics of the secondary school attended should thus be defined as “potential” characteristics (say, X_{i2}^0 and X_{i2}^1), had the child attended a comprehensive (or a selective) school.

⁶In 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).

⁷It can be shown that the distribution of potential outcomes is *not* identified when no restrictions are imposed on the g functions in (1). Even in the case where g is additive in the endowment but depends on unknown parameters as in (2), we will need a third period to ensure identification, see the discussion in Subsection 2.3.

⁸In 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.

⁹We use j instead of i to avoid confusion with the indexation of individual units.

¹⁰Throughout 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.

¹¹Actually, the weaker condition that the real zeros of $\Psi_{Y_{i1}|D_i=0}$ are *isolated* is sufficient for identification. In particular, this result applies to distributions with bounded support, the characteristic functions of which necessarily have isolated real zeros. See Carrasco and Florens (2009) for an approach to identification and estimation that covers the case of isolated real zeros.

¹²In 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: $|\Psi_W(t)| = |\mathbb{E}(\exp(jtW))| \leq \mathbb{E}(|\exp(jtW)|) = 1$.

¹³Remark that 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.

¹⁴Using the expressions of the c.d.f.'s and inverse c.d.f.'s of normal distributions.

¹⁵The assumption that the distributions of unobservables are identical in the two periods is also made in a related paper by Thuysbaert (2007).

¹⁶Another difference with AI is that their identification 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 the estimation section).

¹⁷So, \tilde{Y}_{i0} does *not* satisfy the conditions of an “instrumental variable” in the sense of the literature on heterogeneous treatment effects. Were such an instrumental variable available, distributional treatment effects could be identified on the subpopulation of compliers (e.g., Abadie, Angrist and Imbens, 2002).

¹⁸Identifying restrictions on λ_1 and λ_2^0 may be obtained by taking quasi-differences in (19) and using nonlinear IV (e.g., Amemiya, 1985, p.250).

¹⁹See <http://www.cemfi.es/~bonhomme/>

²⁰One may interpret the set of all pairwise sums $Y_{i2} + Y_{k1}$ such that $D_i = 0$ and $D_k = 1$ as our “sample”, the empirical characteristic function being:

$$\begin{aligned} \frac{1}{N_0 N_1} \sum_{(i,k), D_i=0, D_k=1} \exp[jt(Y_{i2} + Y_{k1})] &= \frac{1}{N_0} \sum_{i, D_i=0} \exp(jtY_{i2}) \cdot \frac{1}{N_1} \sum_{k, D_k=1} \exp(jtY_{k1}) \\ &= \hat{\Psi}_{Y_{i2}|D_i=0}(t) \hat{\Psi}_{Y_{i1}|D_i=1}(t). \end{aligned}$$

²¹In the empirical part we will also report inverse probability weighting estimates, that account for selection on observables only. In that case, we will set ρ to zero in (23).

²²Note that (24) 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 estimated mean. These operations had very little effect on the estimated densities and quantiles.

²³We use the trapezoid rule, with 200 equidistant nodes.

²⁴The asymptotic properties of the c.d.f. and quantile estimators are derived in a recent paper by Hall and Lahiri (2008), in the nonparametric deconvolution problem with known error distribution. In this case also, we are not aware of a formal proof of consistency of the bootstrap.

²⁵In the 1960s and 1970s, the allocation to a specific secondary school was decided at the local level by the LEA. The LEAs had responsibility for most of the spending of secondary schools. See BS for more details.

²⁶When 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.

²⁷An insignificant effect of a comparable magnitude is also obtained when controlling for the first period's test score, see the second row in the table.

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

²⁹In 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.

³⁰That is: rank invariance holds if $F_{Y_{i2}|D_i=1}(Y_{i2}) = F_{Y_{i2}^0|D_i=1}(Y_{i2}^0)$.

³¹Rank 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 $\mathbb{E}(Y_{i2}^0|G_i = 1)$.

³²These 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, although he does find larger differences in longer-run outcomes.

Table 1: Descriptive statistics

Variable	All schools					
	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

Variable	Selective schools					
	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 2: ρ estimates, using various sets of instruments

Instruments		(1)	(2)	(3)	(4)
Mathematics and reading scores (age 7)	ρ 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)	ρ 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	ρ 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. Specification (4) is (3) plus squares and interactions. Standard errors clustered at the LEA level in parentheses.

Table 3: ATT estimates of attending a selective school on the score in mathematics at age 16

Selection on observables					
	(1)	(2)	(3)	(4)	(5)
Inverse probability weighting (IPW)	1.555 (0.267)	1.514 (0.292)	1.449 (0.408)	0.789 (0.701)	1.261 (0.466)
IPW (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					
Instruments	(1)	(2)	(3)	(4)	(5)
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: Estimates of the effect of attending a selective school, average effect on the treated. Specification (1) includes family characteristics, (2), (3) and (4) include in addition school, school and local, and school, local and additional LEA controls, respectively. Specification (5) is (3) plus squares and interactions. Bootstrapped standard errors clustered at the LEA level in parentheses (100 replications).

Table 4: 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)

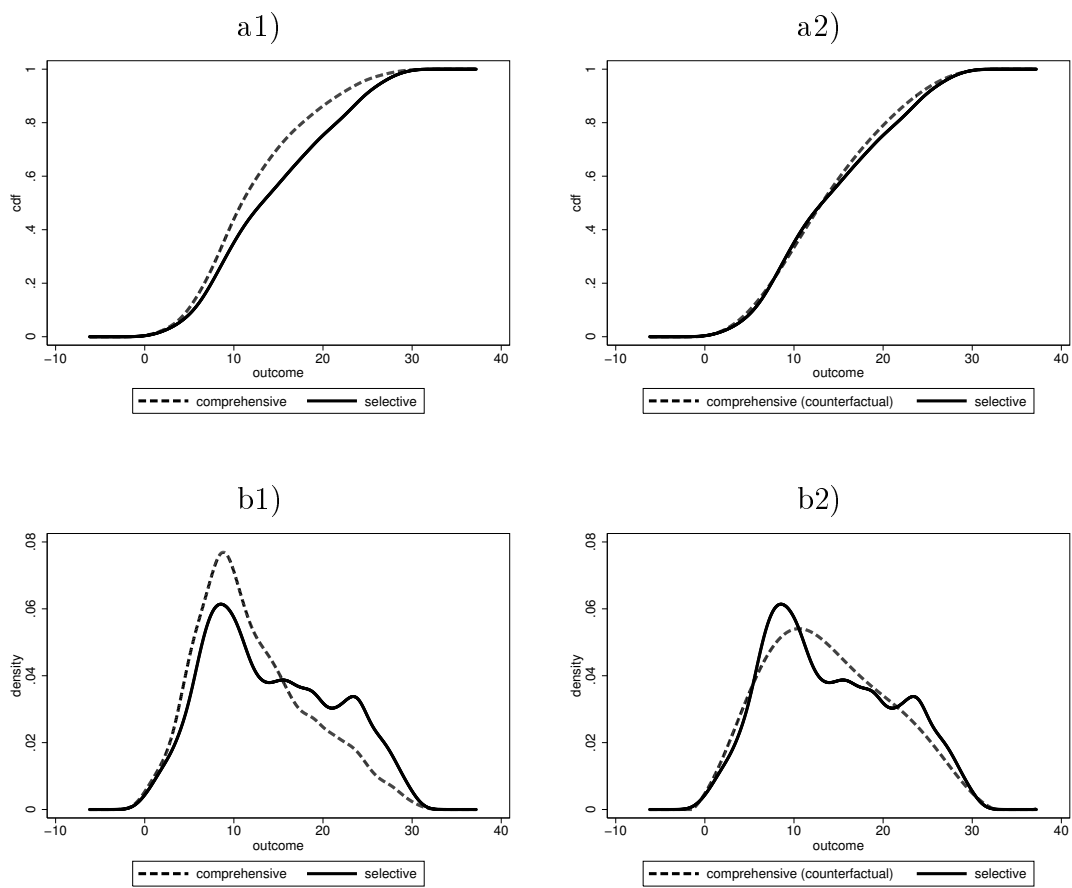
Note: Estimates of the effect of attending a grammar school versus attending a comprehensive school (Δ^G), and of the effect of attending a secondary modern school versus attending a comprehensive school (Δ^S), average effects on the treated. The estimates are recovered from the estimated counterfactual density of outcomes. Covariates specifications are as in Table 3. Bootstrapped standard errors clustered at the LEA level in parentheses (100 replications).

Table 5: 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)

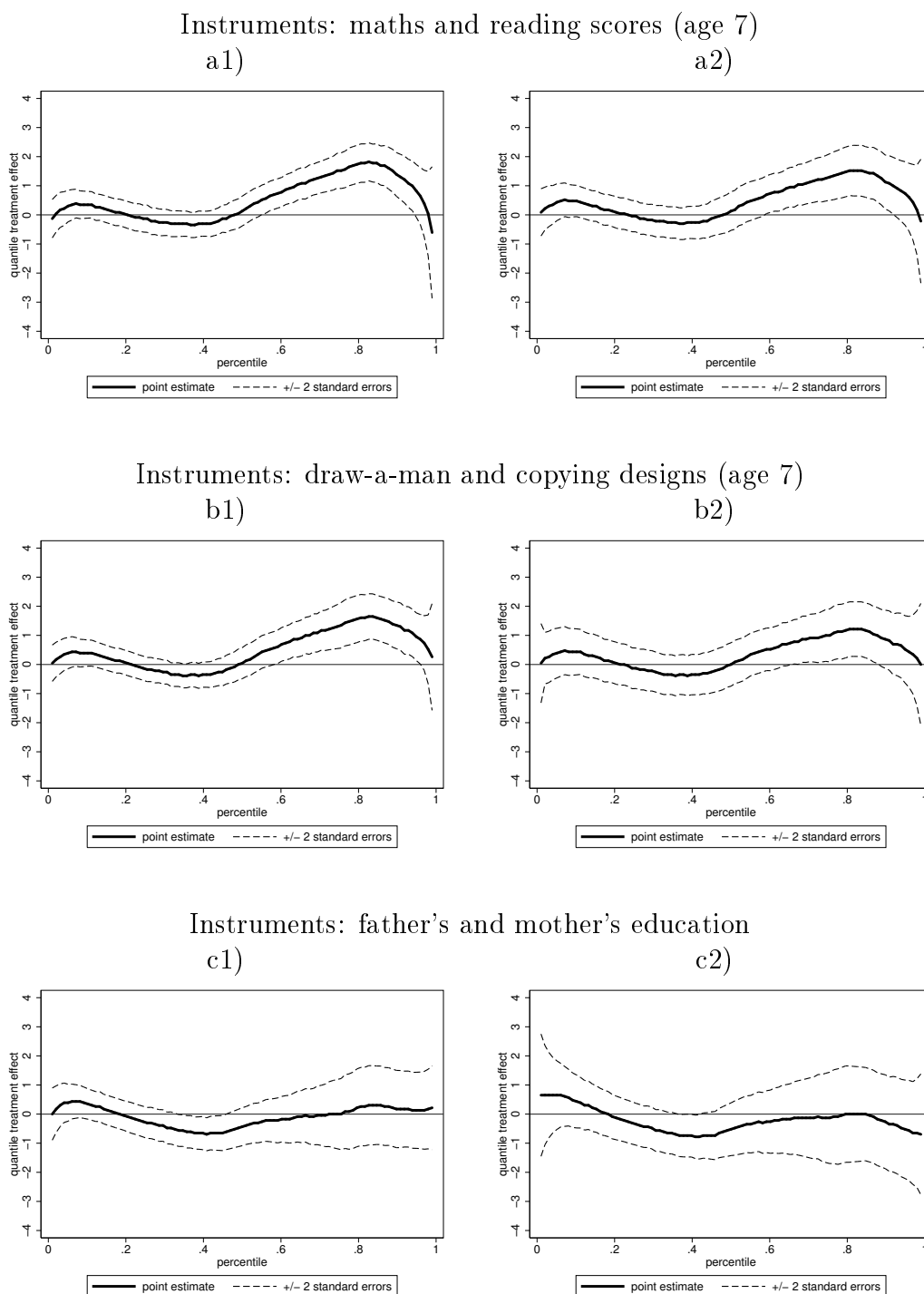
Note: Average treatment effect on the treated of attending a selective rather than a comprehensive school on age 11 test scores. Covariates specification includes family characteristics, as well as age 7 reading and mathematics test scores in the second row. Father's and mother's education are used as 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 3. Draw-a-man and copying designs are used as instruments to estimate ρ . 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 ρ . Solid lines show point estimates, dashed lines show confidence bands of ± 2 bootstrapped standard errors, clustered at the LEA level (100 replications).