

Evaluating the effect of education on earnings: models, methods and results from the National Child Development Survey

Richard Blundell

University College London and Institute for Fiscal Studies, London, UK

and Lorraine Dearden and Barbara Sianesi

Institute for Fiscal Studies, London, UK

[Received June 2001. Final revision September 2004]

Summary. Regression, matching, control function and instrumental variables methods for recovering the effect of education on individual earnings are reviewed for single treatments and sequential multiple treatments with and without heterogeneous returns. The sensitivity of the estimates once applied to a common data set is then explored. We show the importance of correcting for detailed test score and family background differences and of allowing for (observable) heterogeneity in returns. We find an average return of 27% for those completing higher education *versus* anything less. Compared with stopping at 16 years of age without qualifications, we find an average return to O-levels of 18%, to A-levels of 24% and to higher education of 48%.

Keywords: Control function; Evaluation; Heterogeneity; Instrument variables; Matching; Multiple treatments; Non-experimental methods; Propensity score; Returns to education; Selection

1. Introduction

With extensive data available over time and individuals on schooling and on earnings, the measurement of the causal effect of education on earnings is one area where we might expect agreement. However, the literature reveals a wide range of estimates. Many of the differences in estimates reflect genuine differences across the types of educational qualification and the types of individual being analysed. But other differences are a result of the statistical approach that is adopted to recover the effect of education on earnings. The aim of this paper is to provide an empirical and methodological comparison of different approaches. In so doing we provide some new contributions. First, four popular estimation methods—least squares, matching, control function and instrumental variables (IVs)—are compared both from a methodological point of view within a common framework and in terms of the sensitivity of the resulting estimates when applied to a common data set. Secondly, by contrasting the relative magnitude of the various estimates, we try to infer what kind of selection and outcome models underlie our data. The control function provides us with the basis for assessing the importance of residual selection on unobserved returns as well as unobserved individual heterogeneity. We also devote attention to matching, both in its links to simple least squares regression and in terms of the insights

Address for correspondence: Barbara Sianesi, Institute for Fiscal Studies, 7 Ridgmount Street, London, WC1E 7AE, UK.

E-mail: barbara.s@ifis.org.uk

that it can provide in the interpretation of the results. Thirdly, we use the uniquely rich data from the British cohort studies, in particular the National Child Development Survey (NCDS), to assess the importance of test score and family background information in generating reliable estimates. Finally, our focus on heterogeneity is not limited to individual (observed and unobserved) heterogeneity both in characteristics and in returns but explicitly considers treatment heterogeneity in a multiple-treatment framework distinguishing between discrete levels of educational qualifications.

We are not the first to consider these issues. Indeed, there is a growing literature that tries to understand the variety of estimates for the returns to education and to point to the ‘correct’ causal estimate. Card (1999) is the most recent comprehensive study. Here we also draw on the study by Angrist (1998) who compared ordinary least squares (OLS), matching and IV estimators in models with heterogeneous treatment effects. Recently, Heckman and Navarro-Lozano (2004) have compared matching, IV and control functions in the estimation of economic choice models. As to the empirical literature on the effect of education on earnings in the UK, most studies use the repeated cross-section that is available in the Family Expenditure Survey, the General Household Survey or the Labour Force Survey. For example, Gosling *et al.* (2000), Schmitt (1995) and McIntosh (2004) focused on the changing returns over time and could not condition on test score and family background information. Harmon and Walker (1995) exploited the natural experiment of a change in the minimum school-leaving age to circumvent the need to observe ability and family background variables. Dearden (1999a, b) and Blundell *et al.* (2000) both used the British NCDS cohort data, although not focusing on a systematic analysis of the type that is discussed in this paper. Overall, for the UK, most researchers choose to adopt qualification-based measures of educational attainment rather than years of education.

We begin our analysis by using a single-treatment specification focusing on the effect of a specific educational level—such as undertaking higher education (HE). We then consider a multiple-treatment model, which distinguishes the effect of many different education levels, thus allowing the attainment of different educational qualifications to have separate effects on earnings. In general, the multiple-treatment model would seem to be a more attractive framework since we shall typically be interested in a wide range of education levels with potentially very different returns. We also highlight the distinction between heterogeneous and homogeneous returns, i.e. whether the response coefficient on the education variable(s) in the earnings equation is allowed to differ across individuals. Observable heterogeneity is straightforward to account for and in our application we extend the least squares, control function and IV estimators in this direction, thus providing a ‘bridge’ to the matching estimator. By contrast, to allow the heterogeneity to be unobservable to the econometrician, but acted on by individuals, completely changes the interpretation and the properties of many common estimators. In addition, defining which average parameter is of interest becomes crucial. Section 2 outlines in detail our overall modelling framework and the more specialized models that it embeds.

Even where there is agreement on the model specification, alternative statistical methods can be adopted to estimate these models. With experimental data, the standard comparison of a control and treatment group recovers an estimate of the average return for the treated individuals under the assumption that the controls are unaffected by the treatment. Although experimental design is possible and growing in popularity in some studies of training, for large reforms to schooling and for measuring the effect of existing educational systems, non-experimental methods are essential. There are broadly two categories of non-experimental methods: those that attempt to control for the correlation between individual factors and schooling choices by way of an excluded instrument and those that attempt to measure all individual factors that may be the cause of such dependence and then to match on these observed variables. Although the

feasibility of these alternative methods clearly hinges on the nature of the available data, their implementation and properties differ according to whether the model is one of homogeneous or heterogeneous response and whether schooling is represented through a single or a multiple measure. No given non-experimental estimator is uniformly superior to all others; the choice between the various estimation methods should be guided by the postulated model for the outcome and selection processes and the corresponding parameter of interest to be recovered, as well as the richness and nature of the available data in the application at hand.

Our results argue for a cautious approach to the use of any estimator. These results are derived in Section 4 by using the British NCDS data, a rich longitudinal cohort study of all people who were born in Britain in a week in March 1958. In particular, our results show the importance of test score and family background information in pinning down the effect of educational qualifications on earnings. When adopting a matching estimator, they point towards a careful choice of matching variables and highlight the difference in interpretation between measuring the effect of an educational qualification among those who received the qualification and those who did not. They also suggest the use of a control function approach to assess the validity of assumptions on selection where suitable exclusion restrictions can be found.

The estimates point to an average return of about 27% for those completing some form of HE compared with anything less. Compared with leaving school at 16 years of age without qualifications, we also find an average return to O-levels of around 18%, to A-levels of 24% and to HE of 48%. This finding implies that the annualized rate of return is 9.5% compared with leaving school at 16 years of age without qualifications. However, the distinction between those who leave school at the minimum school-leaving age with and without (O-level) qualifications makes it more difficult to estimate a unique rate of return to 'years' of education. In fact, when the base-line comparison is with leaving school at 16 years of age irrespective of qualifications, the average return to a year of HE falls to around 6.6%. If we instead annualize the returns to A-levels (compared with leaving at 16 years of age irrespective of qualifications), the average return per year of education is even lower, at 5.6%. In comparison with US studies, we thus find less evidence of a 'linear' relationship between years of schooling and earnings; educational stages seem to matter.

It may be worth pointing out that, in line with most microeconomic literature, we uniquely focus on the private return to education, ignoring any potential externalities that may benefit the economy at large. In addition, the average individual 'return' to education that we report here is only one component in a full analysis of the private returns to education, which would have to balance individual costs against a flow of such returns over the working life. Moreover, we say nothing about the riskiness of returns to education, which is an important determinant of educational choices among less wealthy families.

The paper proceeds as follows. In Section 2, the single-treatment and multiple-treatment specifications are examined. Section 3 then compares the least squares, matching, control function and IV estimators for these specifications. The estimators are then empirically contrasted in Section 4, focusing on men to avoid confounding issues arising from selection into employment. We first consider a simple single-treatment model looking at the return from undertaking some form of HE (college education). We then move on to look at the estimates from a multiple-treatment model, which includes HE as well as lower level school qualifications and their equivalents. In Section 5, we highlight our main methodological conclusions.

2. The general modelling framework

The problem of measuring the effect of education on earnings falls quite neatly into the evalu-

ation literature: the measurement of the causal impact of a generic ‘treatment’ on an outcome of interest (see, for example, Card (2001) and Heckman *et al.* (1999)). To cover a fairly flexible representation of schooling, we shall consider the *multiple-treatment* case of a finite set of highest schooling levels that are attainable by any given individual. We write the exhaustive set of $J + 1$ treatments (schooling levels) under examination as $0, 1, \dots, J$ and denote the attainment by individual i of schooling level j as his or her highest level by $S_{ji} = 1$. This specification is very flexible and can cover education outcomes that occur in some natural sequence—including completion of j years of schooling by individual i .

We can think of a set of potential outcomes associated to each of the $J + 1$ treatments: $y_i^0, y_i^1, \dots, y_i^J$, where y_i^j denotes the (log-) earnings of individual i if i were to receive schooling level j . The problem of estimating the returns to education can be phrased as the evaluation of the causal effect of one schooling level j relative to another (without loss of generality, let this be treatment 0) on the outcome considered, y . In terms of the notation that we have established, interest will lie in recovering quantities of the form $y_i^j - y_i^0$, averaged over some population of interest, such as the whole population or those who did achieve that level.

Each individual, however, receives only one of the treatments, and the remaining J potential outcomes are unobserved counterfactuals. At the core of the evaluation problem, including its application to the returns-to-education framework, is thus the attempt to estimate missing data. The observed outcome of individual i can be written as

$$y_i = y_i^0 + \sum_{j=1}^J (y_i^j - y_i^0) S_{ji}. \tag{1}$$

Equation (1) is extremely general; however, we require some further notation before we can discuss the various models and estimation methods that are the subject of this paper. We let potential outcomes depend on both observed covariates X_i and unobserved factors u_i^j in the following general way:

$$y_i^j = f_j(X_i, u_i^j) \quad \text{for } j = 0, 1, \dots, J. \tag{2}$$

For this representation to be meaningful, the stable unit-treatment value assumption needs to be satisfied (Rubin (1980), and, for further discussions, Rubin (1986) and Holland (1986)). This assumption requires that an individual’s potential outcomes as well as the chosen education level are independent from the choices of schooling of other individuals in the population, thus ruling out spill-over or general equilibrium effects. Note also that implicit in expression (2) is the requirement that the observables X be exogenous in the sense that their potential values do not depend on treatment status, or, equivalently, that their potential values for the different treatment states coincide ($X_{ji} = X_i$ for $j = 0, 1, \dots, J$). Natural candidates for X that are not determined or affected by treatments S are time constant factors, as well as pretreatment characteristics.

Assuming additive separability between observables and unobservables, we can write

$$y_i^j = m_j(X_i) + u_i^j$$

with $E[y_i^j | X_i] = m_j(X_i)$, i.e. assuming that the observable regressors X are unrelated to the unobservables u . We shall maintain these exogeneity assumptions on the X s throughout.

Let the state-specific unobservable components of earnings be written as

$$u_i^j = \alpha_i + \varepsilon_i + b_{ji} \quad \text{for } j = 0, 1, \dots, J$$

with α_i representing some unobservable individual trait, such as ability or motivation, that affects earnings for any given level of schooling, b_{ji} measuring the individual-specific unobserved

marginal return to schooling level j relative to level 0 in terms of the particular definition of earnings y_i (for convenience, let us normalize b_{0i} to 0) and ε_i being the standard residual, possibly capturing measurement error in earnings as well (measurement error in the schooling variable S may also be important and will be touched on later).

Given this general specification, equation (1) for observed earnings becomes

$$\begin{aligned}
 y_i &= m_0(X_i) + \sum_{j=1}^J \{m_j(X_i) - m_0(X_i)\} S_{ji} + \sum_{j=1}^J (u_i^j - u_i^0) S_{ji} + \alpha_i + \varepsilon_i \\
 &= m_0(X_i) + \sum_{j=1}^J b_j(X_i) S_{ji} + \sum_{j=1}^J b_{ji} S_{ji} + \alpha_i + \varepsilon_i \\
 &= m_0(X_i) + \sum_{j=1}^J \beta_{ji} S_{ji} + \alpha_i + \varepsilon_i
 \end{aligned} \tag{3}$$

with $\beta_{ji} \equiv b_j(X_i) + b_{ji}$.

In this set-up, β_{ji} , the private return to schooling level j (relative to schooling level 0), is allowed to be heterogeneous across individuals in both observable and unobservable dimensions; $b_j(X_i)$ represents the return for individuals with characteristics X_i and thus captures observable heterogeneity in returns, whereas b_{ji} represents the individual-specific unobserved return to schooling level j , conditional on X_i . Typically, we would assume the α_i and b_{ji} to have a finite population mean (denoted by α_0 and b_{j0} respectively) and variance.

With this general specification in place, we can now look at the differences between the homogeneous and heterogeneous returns models and within these models look at differences between single-treatment, multiple-treatment and one-factor models.

2.1. The homogeneous returns model

In the homogeneous returns framework, the rate of return to a given schooling level j is the same across individuals, i.e. $\beta_{ji} = \beta_j$ for all individuals i . In the case of a finite set of schooling levels (specific discrete educational levels as in the application that will be used in our paper, or even finer with each level representing a year of education), the *multiple-treatment model* (3) becomes

$$y_i = m_0(X_i) + \beta_1 S_{1i} + \beta_2 S_{2i} + \dots + \beta_J S_{Ji} + \alpha_i + \varepsilon_i \tag{4}$$

where α_i represents differing relative levels of earnings across individuals for any given level of schooling and the β_j s measure the effect of schooling level j relative to the base level. Although the returns to a given level are homogeneous across individuals, the different schooling levels are allowed to have different effects on earnings.

This is not true in the popular *one-factor human capital model*, where it is assumed that education can always be aggregated into a single measure, say years of schooling, $S_i \in \{0, 1, \dots, J\}$. In this specification,

$$y_i = m_0(X_i) + \beta S_i + \alpha_i + \varepsilon_i \tag{5}$$

which can be obtained from our general set-up (3) with the various treatment levels as years of education (so that $S_i = \sum_{j=1}^J j S_{ji}$ with $S_{ji} \equiv \mathbf{1}_{(S_i=j)}$) by assuming the linear relationship $\beta_{ji} = \beta_j = j\beta$ —i.e. that the (homogeneous) return to j years of schooling is simply j times the return to one year of schooling—or, equivalently, $\beta_{j+1,i} - \beta_{ji} = \beta$ for all $j = 0, 1, \dots, J$ —i.e. that each additional year of schooling has the same marginal return.

A final specification, which can be obtained from model (3) by setting $J = 1$, is the *single-treatment model*, the aim of which is to recover the causal impact of a single type of schooling level $S_1 \in \{0, 1\}$ —e.g. undertaking HE or college compared with not doing so. In the homogeneous returns model, this single-treatment specification can be expressed as

$$y_i = m_0(X_i) + \beta S_{1i} + \alpha_i + \varepsilon_i$$

where β is the return to achieving the education level under consideration (relative to educational level 0 as chosen for $S_{1i} = 0$).

Although in these homogeneous returns models β_{ji} is constant across all individuals, α_i is allowed to vary across i to capture the differing productivities (or abilities or earnings levels) across individuals with the same education levels. Since educational choices and thus attained educational levels are likely to differ according to productivity (or expected earnings levels more generally), the schooling variable S is very likely to be correlated with α_i and this in turn will induce a bias in the simple least squares estimation of β . In addition, if S is measured with error, there will be some further bias. We shall return to these estimation issues in more detail later.

2.2. The heterogeneous returns model

Despite the preponderance of the homogeneous returns model in the early literature, the recent focus has been on models allowing for heterogeneous returns (examples include Card (2001), Heckman, Smith and Clements (1997), Dearden (1999a, b) and Blundell *et al.* (2000)). Once the return is allowed to vary across individuals, the immediate question concerns the parameter of interest. Is it the average of the individual returns? If so, what average is it? Is it the average in the population whether or not the educational level under consideration is achieved—the *average treatment effect*, ATE—or the average among those individuals actually observed to achieve the educational level—the *average effect of treatment on the treated*, ATT—or the average among those who have not achieved that educational level—the *average effect of treatment on the non-treated*, ATNT? In some cases, a change in policy can be used to recover a *local average treatment effect* (LATE), measuring the return for an even smaller subgroup of individuals: those induced to take the educational level by the change in policy. We discuss all these in greater detail in the next section.

In the general framework (3), the return to schooling level j is allowed to be heterogeneous across individuals in both observable and unobservable dimensions. It is straightforward to generalize models such as equations (4) or (5) to allow for the *observable heterogeneity* $b_j(X_i)$. What is more difficult is how we deal with *unobserved heterogeneity* across individuals in the response parameter β . This person-specific component of the return may be observed by the individual but is unobserved by the analyst.

Consider first the single-treatment model. A general relationship between the level of education under examination and earnings is then written as

$$\begin{aligned} y_i &= m_0(X_i) + \beta_i S_{1i} + \alpha_i + \varepsilon_i \\ &= m_0(X_i) + \{b(X_i) + b_0\} S_{1i} + (b_i - b_0) S_{1i} + \alpha_i + \varepsilon_i \end{aligned} \quad (6)$$

where b_i can be thought of as random coefficients representing the heterogeneous relationship between educational qualification S_{1i} and earnings, conditional on observables X_i ($b_0 \equiv E[b_i]$ denoting its population mean).

The parameter of interest will be some average of $b(X_i) + b_i$, with the average taken over the subpopulation of interest; the resulting parameter will thus measure the average return

to achieving education level S_1 for this group. Examples are the average effect of treatment on the treated, $\beta_{ATT} \equiv E[b(X_i) + b_i | S_{1i} = 1]$, the average treatment effect in the population, $\beta_{ATE} \equiv E[b(X_i)] + b_0$, and the average effect on the non-treated, $\beta_{ATNT} \equiv E[b(X_i) + b_i | S_{1i} = 0]$.

As we mentioned in the homogeneous models above, the dependence of the schooling level(s) on the unobserved ‘ability’ component α_i is critical in understanding the bias from the direct comparison of groups with and without education level S_1 . A further key issue in determining the properties of standard econometric estimators in the heterogeneous effects model is whether or not schooling choices S_{1i} depend on the unobservable determinants of the individual’s marginal return from schooling b_i , conditional on observables X_i . If, given the information in X_i , there is some gain b_i that is still unobserved by the econometrician but known in advance (or predictable) by the individual when making his or her educational choices, then it would seem sensible to assume that choices will, in part at least, reflect the return to that choice. Since, however, b_i is likely to vary over time and will depend on the relative levels of demand and supply, the dependence of choices of schooling on marginal returns is not clear cut. Some persistence in returns is, however, likely, and so some correlation would seem plausible.

The discussion of heterogeneous returns extends easily to the multiple-treatment model (4):

$$y_i = m_0(X_i) + \beta_{1i}S_{1i} + \beta_{2i}S_{2i} + \dots + \beta_{Ji}S_{Ji} + \alpha_i + \varepsilon_i. \tag{7}$$

In fact, the three basic specifications (6), (4) and (7) will form the main alternatives that are considered in the paper, the single discrete treatment case (6) being the base-line specification.

3. Estimation methods

The aim of this section is to investigate the properties of alternative non-experimental estimation methods for each of the model specifications that were considered above. We begin by considering a naïve estimator in the general framework (3) of the returns to educational level j (relative to level 0) for individuals reaching this level: the simple difference between the observed average earnings of individuals with $S_{ji} = 1$ and the observed average earnings of individuals with $S_{0i} = 1$.

This observed difference in conditional means can be rewritten in terms of the average effect of treatment on the treated parameter (what we are after) and the bias that potentially arises when the earnings of the observed group with $S_{0i} = 1$ ($y_i^0 | S_{0i} = 1$) are used to represent the counterfactual ($y_i^0 | S_{ji} = 1$):

$$\begin{aligned} \text{naïve estimator} &\equiv E[y_i | S_{ji} = 1] - E[y_i | S_{0i} = 1] \\ &= E[y_i^j - y_i^0 | S_{ji} = 1] - (E[y_i^0 | S_{ji} = 1] - E[y_i^0 | S_{0i} = 1]) \\ &= \text{ATT} - \text{bias}. \end{aligned}$$

The key issue is that, since educational choices are likely to be the result of systematic decisions, the sample of individuals who make each choice will not be random. If this is ignored and individuals who make the choice are simply compared with those who did not, the estimates will suffer from bias.

Using experimental data, Heckman, Ichimura, Smith and Todd (1998) provided a very useful breakdown of this bias term:

$$\text{bias} \equiv E[y_i^0 | S_{ji} = 1] - E[y_i^0 | S_{0i} = 1] = B_1 + B_2 + B_3. \tag{8}$$

The first two components in expression (8) arise from differences in the distribution of observed characteristics X between the two groups: B_1 represents the bias component due to

non-overlapping support of the observables and B_2 is the error part due to misweighting on the common support, as the resulting empirical distributions of observables are not necessarily the same even when restricted to the same support. The last component, B_3 , is the true econometric selection bias resulting from ‘selection on unobservables’—in our notation, α_i , b_{ji} and ε_i .

Of course, a properly designed randomized experiment would eliminate the bias that was discussed above, but pure education or schooling experiments are very rare. We must instead rely on non-experimental methods, each of which uses observed data together with some appropriate identifying assumptions to recover the missing counterfactual. Depending on the richness and nature of the available data and the postulated model for the outcome and selection processes, the researcher can thus choose from the alternative methods the one that is most likely to avoid or correct the sources of bias that were outlined above. We now look at these methods in turn. The initial setting for the discussion of the three broad classes of alternative methods that we consider—IV, control function and matching—will be based on biases that occur from the simple application of OLS to the various models that were described in the previous section.

3.1. Least squares

Consider the single-treatment model examining the effect of a given educational level S_1 . The model is to be estimated for a given population (defined, for instance, as all those individuals entering schooling at a particular date). In the heterogeneous case, specification (6) is

$$y_i = m_0(X_i) + b(X_i)S_{1i} + b_i S_{1i} + \alpha_i + \varepsilon_i.$$

There are several potential sources of bias in the least squares regression of log-earnings on schooling to recover average treatment effects. The following borrows from the bias decomposition that is highlighted in expression (8).

3.1.1. Bias due to observables: misspecification

First, note that, to implement model (6) parametrically, the functional forms for both

- (a) $E[y_i^0 | X_i] \equiv m_0(X_i)$ and
- (b) $E[y_i^1 - y_i^0 | X_i] \equiv b(X_i)$ need to be specified.

A standard least squares specification would generally control linearly for the set of observables $\{S_{1i}, X_i \equiv \{X_{1i} \dots X_{Mi}\}'\}$ —i.e. it would be of the form

$$y_i = \gamma' X_i + b S_{1i} + \eta_i$$

and thus suffer from two potential sources of bias from observables:

- (a) *misspecification of the no-treatment outcome* $m_0(X_i)$ —if the true model contains higher order terms of the X s, or interactions between the various X s, the OLS estimate of b would in general be biased owing to omitted variables;
- (b) *heterogeneous returns* $b(X_i)$ —simple OLS constrains the returns to be homogeneous. If, by contrast, the effect of schooling varies according to some of the X s, the OLS estimate of b will not in general recover ATT. To illustrate this point, focus on one X -variable and suppose that the true model is $y_i = a + b_0 S_{1i} + b_x X_i S_{1i} + d X_i + e_i$. If we estimate the simple model above, which ignores $X_i S_{1i}$, the estimated coefficient on S_{1i} will have expectation $E[\hat{b}] = b_0 + b_x \phi$, with ϕ defined in $X_i S_{1i} = \pi + \tau X_i + \phi S_{1i} + \nu_i$. OLS will not in general recover $ATT \equiv E[Y^1 - Y^0 | S_1 = 1] = b_0 + b_x E[X | S_1 = 1]$, since in general ϕ is different from $E[X | S_1 = 1]$:

$$\phi = E[X|S_1 = 1] \frac{V(X) - \text{cov}(X, S_1)}{V(X) - \text{cov}(X, S_1)^2 V(S_1)^{-1}}.$$

These misspecification issues are linked to the source of bias B_2 —not appropriately re-weighting the observations to control fully for the difference in the distribution of X over the common region—as well as to source B_1 —a lack of sufficient overlap in the two groups’ densities of X . The OLS approximation of the regression function $m_0(X_i)$ over the non-overlapping region is purely based on the chosen (in our example, linear) functional form; in other words, for treated individuals outside the common support, the OLS identification of the counterfactual crucially relies on being based on the correctly specified model.

3.1.2. *Bias due to unobservables*

Gathering the unobservables together in equation (6), we have

$$y_i = m_0(X_i) + \beta_{\text{ATE}}(X_i)S_{1i} + e_i \quad \text{with } e_i \equiv \alpha_i + (b_i - b_0)S_{1i} + \varepsilon_i, \quad (9)$$

$$y_i = m_0(X_i) + \beta_{\text{ATT}}(X_i)S_{1i} + w_i \quad \text{with } w_i \equiv \alpha_i + (b_i - E[b_i|X_i, S_{1i} = 1])S_{1i} + \varepsilon_i \quad (10)$$

where $\beta_{\text{ATE}}(X_i) \equiv b(X_i) + b_0$ and $\beta_{\text{ATT}}(X_i) \equiv b(X_i) + E[b_i|X_i, S_{1i} = 1]$.

Running a correctly specified OLS regression will produce a biased estimator of either parameter of interest if there is correlation between S_{1i} and the error term e_i or w_i , i.e. $E[e_i|X_i, S_{1i}]$ and $E[w_i|X_i, S_{1i}]$ may be non-zero. Such correlation may arise from different sources.

- (a) *Ability bias* arises from the likely correlation between the α_i intercept term (the absolute advantage) and S_{1i} . If higher ability or inherently more productive individuals tend to acquire more education, the two terms will be positively correlated, inducing an upward bias in the estimated average return β_{ATE} or β_{ATT} .
- (b) *Returns bias* occurs when the individual returns component b_i (the comparative advantage) is itself correlated with the schooling decision S_{1i} . The direction of this bias is less clear and will depend on the average returns among the subpopulation of those with schooling level $S_{1i} = 1$. Indeed, if
 - (i) ability bias is negligible (i.e. $E[\alpha_i|X_i, S_{1i}] = 0$),
 - (ii) the ability heterogeneity is unrelated to the unobserved return and
 - (iii) the returns bias is the only remaining bias present (i.e. $E[b_i|X_i, S_{1i} = 1] \neq b_0$),
 then expressions (9) and (10) show how the least squares coefficient on S_{1i} will be biased for the average treatment effect β_{ATE} but will recover the average effect of treatment on the treated β_{ATT} .
- (c) For *measurement error bias* we can think of ε_i as including measurement error in the schooling variable S_{1i} . Since the educational variable is a dummy or categorical variable, measurement error will be non-classical (in particular, it will vary with the level of education that is reported). Kane *et al.* (1999) showed that both OLS and IV estimates may be biased and that it is not possible to place any *a priori* general restrictions on the direction or magnitude of the bias of either estimator. By contrast, in the case of a continuous variable which is affected by (classical) measurement error, OLS estimates of the return would be downward biased and IV estimates consistent.

In the homogeneous returns model, the second source of bias is, by definition, absent. This is the case that is much discussed in the literature (especially in the one-factor ‘years of schooling’ model (5)), where the upward ability bias may be partially offset by the attenuation measurement error bias; this trade-off was at the heart of the early studies on measuring gross private returns (for a review see in particular Griliches (1977) and Card (1999)).

Much of the practical discussion of the properties of least squares bias depends on the richness of other control variables that may be entered to capture the factors omitted. Indeed, the method of matching takes this further by trying to control directly and flexibly for all those variables at the root of selection bias.

3.2. Matching methods

The general matching method is a nonparametric approach to the problem of identifying the treatment impact on outcomes. To recover the average treatment effect on the treated individuals, the matching method tries to mimic *ex post* an experiment by choosing a comparison group from among the non-treated individuals such that the selected group is as similar as possible to the treatment group in terms of their observable characteristics. Under the matching assumption, all the outcome relevant differences between treated and non-treated individuals are captured in their observable attributes, the only remaining difference between the two groups being their treatment status. In this case, the average outcome of the matched non-treated individuals constitutes the correct sample counterpart for the missing information on the outcomes that the treated would have experienced, on average, if they had not been treated.

The central issue in the matching method is choosing the appropriate matching variables. This is a knife-edge decision as there can be too many as well as too few to satisfy the identifying assumption for recovering a consistent estimate of the treatment effect. In some ways, this mirrors the issue of choosing an appropriate excluded instrument in the IV and control function approaches that are discussed later. However, instruments do not make appropriate matching variables and vice versa. Instruments should satisfy an exclusion condition in the outcome equation conditional on the treatment, whereas matching variables should affect both the outcome and the treatment equations.

3.2.1. General matching methods

To illustrate the matching solution for the average effect of treatment on the treated individuals in a more formal way, consider the *completely general* specification of the earnings outcomes (2)—i.e. the one not even requiring additive separability—in the single discrete treatment case ($J = 1$). Among the set of variables X in the earnings equations, we distinguish those affecting *both* potential no-treatment outcomes y^0 and schooling choices S from those affecting outcomes y^0 alone. We denote the former subset of X by \tilde{X} .

The solution to the missing counterfactual advanced by matching is based on a fundamental assumption of conditional independence between non-treatment outcomes and the schooling variable S_{1i} (matching method assumption 1, MM1):

$$y_i^0 \perp S_{1i} | \tilde{X}_i.$$

This assumption of *selection on observables* requires that, conditionally on an appropriate set of observed attributes, the distribution of the (counterfactual) outcome y^0 in the treated group is the same as the (observed) distribution of y^0 in the non-treated group. For each treated observation ($y_i : i \in \{S_{1i} = 1\}$), we can look for a non-treated (set of) observation(s) ($y_i : i \in \{S_{1i} = 0\}$) with the same \tilde{X} -realization. Under the matching assumption that the chosen group of matched comparisons (i.e. conditionally on the \tilde{X} s that are used to select them) does not differ from the treatment group by any variable that is systematically linked to the non-participation outcome y^0 , this matched comparison group constitutes the counterfactual required.

As should be clear, the matching method avoids defining a specific form for the outcome equation, decision process or either unobservable term. Still, translated into the more specialized

framework of equation (6), MM1 becomes $(\alpha_i, \varepsilon_i) \perp S_{1i} | \tilde{X}_i$. Note that the individual-specific return to education b_i is allowed to be correlated with the schooling decision S_{1i} , provided that in this case $(\alpha_i, \varepsilon_i) \perp b_i | \tilde{X}_i$ also holds. In particular, individuals may decide to acquire schooling on the basis of their individual gain from it (which is unobserved by the analyst), as long as this individual gain is not correlated to their non-treatment outcome y_i^0 conditional on \tilde{X} .

For the matching procedure to have empirical content, it is also required that

$$P(S_{1i} = 1 | \tilde{X}) < 1$$

(matching method assumption 2, MM2), which prevents \tilde{X} from being a perfect predictor of treatment status, guaranteeing that all treated individuals have a counterpart in the non-treated population for the set of \tilde{X} -values over which we seek to make a comparison. Depending on the sample that is in use, this can be quite a strong requirement (e.g. when the education level under consideration is directed to a well-specified group). If there are regions where the support of \tilde{X} does not overlap for the treated and non-treated groups, matching must in fact be performed over the common support region C^* ; the estimated treatment effect must then be redefined as the mean treatment effect for those treated falling within the common support.

To identify the average treatment effect on the treated individuals over C^* , this weaker version in terms of conditional mean independence, which is implied by MM1 and MM2, would actually suffice:

$$E[y^0 | \tilde{X}, S_1 = 1] = E[y^0 | \tilde{X}, S_1 = 0] \quad \text{for } \tilde{X} \in C^*.$$

(matching method assumption 1', MM1').

On the basis of these conditions, a subset of comparable observations is formed from the original sample, and with those a consistent estimator for the treatment effect on the treated individuals (within the common support C^*) is, simply, the mean conditional difference in earnings over C^* , appropriately weighted by the distribution of \tilde{X} in the treated group.

The preceding discussion has referred to the estimation of the average treatment effect on the treated group. If we are also interested in using matching to recover an estimate of the effect of treatment on the non-treated group, as we do in our application to the NCDS data, a symmetric procedure applies, where MM2 needs to be extended to $0 < P(S_{1i} = 1 | \tilde{X}_i)$ for $\tilde{X} \in C^*$ and MM1 to include y^1 . In terms of the framework of equation (6), the strengthened MM1 thus becomes $(\alpha_i, \varepsilon_i, b_i) \perp S_{1i} | \tilde{X}_i$, highlighting how now possibly heterogeneous returns b_i are prevented from affecting educational choices by observably identical agents. Under these strengthened assumptions, the average treatment effect $E[y^1 - y^0]$ can then be simply calculated as a weighted average of the effect on the treated and the effect on the non-treated individuals.

As to the potential sources of bias that were highlighted by the decomposition in expression (8), matching corrects for the first two, B_1 and B_2 , through the process of choosing and re-weighting observations within the common support. In fact, in the general nonparametric matching method, quite a general form of $m_0(X)$ and of interactions $b(X)S_{1i}$ is allowed (note the use of X rather than \tilde{X} —matching would balance also the variables affecting outcomes alone, since by construction they would not differ between treatment groups), avoiding the potential misspecification bias that was highlighted for OLS. Arguing the importance of the remaining source of bias—the one due to unobservables—amounts to arguing the inadequacy of the conditional independence assumption (MM1) in the specific problem at hand, which should be done in relation to the richness of the available observables (i.e. the data \tilde{X}) in connection with the selection and outcome processes.

Turning now to the implementation of matching estimators, consider ATT (similar procedures obviously apply to ATNT). On the basis of MM1' but without invoking any functional form assumption, ATT can be estimated by performing any type of nonparametric estimation of the conditional expectation function in the non-treated group, $E[y_i | S_{1i} = 0, \tilde{X}]$, and averaging it over the distribution of \tilde{X} in the treated group (within the common support). Matching, like for instance stratification on \tilde{X} , is one possible way of performing such a nonparametric regression. The main idea of matching is to pair to each treated individual i some group of 'comparable' non-treated individuals and then to associate to the outcome y_i of treated i a matched outcome \hat{y}_i given by the (weighted) outcomes of his or her 'neighbours' in the comparison group.

The general form of the matching estimator for the average effect of treatment on the treated group (within the common support) is then given by

$$\hat{\beta}_{MM} = \hat{\beta}_{ATT} = \frac{1}{N_1^*} \sum_{i \in \{S_{1i}=1 \cap C^*\}} (y_i - \hat{y}_i)$$

where N_1^* is the number of treated individuals falling within the common support C^* . In particular, $(1/N_1^*) \sum \hat{y}_i$ is the estimate of the average no-treatment counterfactual for the treated individuals, $E[y^0 | S_1 = 1]$.

The general form for the outcome to be paired to treated i 's outcome is

$$\hat{y}_i = \sum_{j \in C^0(\tilde{X}_i)} W_{ij} y_j \tag{11}$$

where

- (a) $C^0(\tilde{X}_i)$ defines treated observation i 's neighbours in the comparison group (where proximity is in terms of their characteristics to i 's characteristics \tilde{X}_i) and
- (b) W_{ij} is the weight that is placed on non-treated observation j in forming a comparison with treated observation i ($W_{ij} \in [0, 1]$ with $\sum_{j \in C^0(\tilde{X}_i)} W_{ij} = 1$).

Although the various matching estimators are all consistent, in finite samples they may produce different estimates as they differ in the way that they construct the matched outcome \hat{y} . Specifically, differences will depend on how they define the neighbourhood in the non-treated group for each treated observation, and, related to this, in how they choose the weights.

The traditional and most intuitive form of matching is *nearest neighbour* (or *one-to-one*) matching, which associates to the outcome of treated unit i a 'matched' outcome given by the outcome of the most observably similar non-treated unit. A variant of nearest neighbour matching is *caliper matching* (see Cochran and Rubin (1973) and, for a recent application, Dehejia and Wahba (1999)). The 'caliper' is used to exclude observations for which there is no close match, thus enforcing common support. A different class of matching estimators has recently been proposed by Heckman, Ichimura and Todd (1997, 1998) and Heckman, Ichimura, Smith and Todd (1998). In *kernel-based matching*, the outcome y_i of treated unit i is matched to a weighted average of the outcomes of more (possibly all) non-treated units, where the weight that is given to non-treated unit j is in proportion to the closeness of the characteristics of i and j . The weight in equation (11) above is set to

$$W_{ij} = K\left(\frac{\tilde{X}_i - \tilde{X}_j}{h}\right) / \sum_{j \in C^0(\tilde{X}_i)} K\left(\frac{\tilde{X}_i - \tilde{X}_j}{h}\right),$$

where $K(\cdot)$ is a non-negative, symmetric and unimodal function, such as the Gaussian kernel $K(u) \propto \exp(-u^2/2)$ or the Epanechnikov kernel $K(u) \propto (1 - u^2) \mathbf{1}(|u| < 1)$.

3.2.2. High dimensionality and the propensity score

It is clear that, when a wide range of \tilde{X} -variables is in use, finding exact matches can be extremely difficult. One possibility to reduce the high dimensionality of the problem is to use some metric to combine all the matching variables into a scalar measuring the distance between any two observations. An attractive, unit-free metric is the Mahalanobis metric, which assigns weight to each co-ordinate of \tilde{X} in proportion to the inverse of the variance of that co-ordinate. The distance between observations i and j is thus defined as $d(i, j) = (\tilde{X}_i - \tilde{X}_j)' V^{-1} (\tilde{X}_i - \tilde{X}_j)$, with V being the covariance matrix of \tilde{X} in the sample (see Abadie and Imbens (2002) and Zhao (2004) for alternative matching metrics).

Following Rosenbaum and Rubin (1983), distance can also be measured in terms of a *balancing score* $q(\tilde{X})$, which is defined as a function of the observables such that $\tilde{X} \perp S_1 | q(\tilde{X})$. One such balancing score is the *propensity score*, the probability to receive treatment given the set of observed characteristics jointly affecting treatment status and outcomes: $p(\tilde{X}_i) \equiv P(S_{1i} = 1 | \tilde{X}_i)$. By definition, treatment and non-treatment observations with the same value of the propensity score have the same distribution of the full vector of regressors \tilde{X} . Rosenbaum and Rubin (1983) have further shown that, under MM1 and MM2 (i.e. when $(y^1, y^0) \perp S_1 | \tilde{X}$ and $0 < p(\tilde{X}) < 1$), then $(y^1, y^0) \perp S_1 | p(\tilde{X})$. In other words, the conditional independence assumption remains valid if $p(\tilde{X})$ —a scalar variable—is used for matching rather than the complete vector of \tilde{X} .

Propensity score matching thus reduces the high dimensional nonparametric estimation problem to a one-dimensional problem: the estimation of the mean outcome in the non-treated group as a function of the propensity score. Again, there are various ways to perform this one-dimensional nonparametric regression, such as stratification on the propensity score, weighting on the propensity score or the matching estimators that were outlined above, where in the formulae for W_{ij} and $C^0(\cdot)$ the vector \tilde{X}_i is simply replaced by the scalar $p_i \equiv p(\tilde{X}_i)$.

It should, however, be noted that in empirical applications the propensity score first needs to be estimated. Since a fully nonparametric estimation of the propensity score would be liable to suffer from the same curse of dimensionality as the standard matching estimator, the estimation task is generally accomplished parametrically, e.g. via a logit or probit specification. The validity of the chosen specification for the propensity score can then be tested against a nonparametric alternative. Propensity score matching thus becomes a *semiparametric* approach to the evaluation problem (see Imbens (2004) for a review of fully nonparametric estimators based on MM1). The estimated propensity score is used only in a first step to correct (parametrically) for the selection bias (on observables) by selecting that subset of the non-treated group to act as a comparison group or, more generally, by appropriately reweighting the non-treated group. All that is required is in fact its ability to balance the relevant observables in the two matched groups ($\tilde{X} \perp S_1 | \hat{p}(\tilde{X})$). Simple parametric specifications for the propensity score have indeed often been found to be quite effective in achieving the balancing required (see for example Zhao (2004)). The second step, the estimation of the treatment effect, can then be accomplished in a fully nonparametric way, in particular without imposing any functional form restriction on how the treatment effect or the no-treatment outcome can vary according to \tilde{X} . The curse of dimensionality is thus side-stepped by parametrically estimating the propensity score only, whereas the specification of $E[y^1 - y^0 | X]$ and of $E[y^0 | X]$ is left completely unrestricted.

The estimation of the standard errors of the treatment effects should ideally adjust for the additional sources of variability that are introduced by the estimation of the propensity score as well as by the matching process itself. For kernel-based matching, analytical asymptotic results have been derived by Heckman, Ichimura and Todd (1998), whereas for one-to-one matching

the common solution is to resort to bootstrapped confidence intervals. (For a comparison of the small sample properties of matching estimators, see Abadie and Imbens (2002), Angrist and Han (2004), Frölich (2004) and Zhao (2004).)

Before concluding this overview of the implementation of propensity score matching estimators, we briefly consider how the various types implement the common support requirement. Simple nearest neighbour matching does not impose any *a priori* common support restriction. In fact, the nearest neighbour could at times turn out to be quite apart. By contrast, its caliper variant, provided that it is not too 'tolerant' (as perceived by the researcher), automatically uses the observations within the common support of the propensity score. As to kernel-based matching estimators, two factors automatically affect the imposition of common support: the choice of bandwidth (a small bandwidth amounts to being very strict in terms of the distance between a non-treated unit and the treated unit under consideration, *de facto* using—i.e. placing weight on—only those comparisons in a close neighbourhood of the treated unit's propensity score) and, to a lesser extent, the choice of kernel (for example, to smooth at a given p_i , the Gaussian kernel uses all the non-treated units, i.e. $C^0(p_i) = \{j: S_{1j} = 0\}$, whereas the Epanechnikov kernel uses only those non-treated units whose propensity score falls within a fixed radius h from p_i , i.e. $C^0(p_i) = \{j \in \{S_{1j} = 0\} : |p_i - p_j| < h\}$). Typically in kernel-based matching the common support is additionally imposed on treated individuals at the boundaries: those treated whose propensity score is larger than the largest propensity score in the non-treated pool are left unmatched. A more refined procedure was suggested by Heckman, Ichimura and Todd (1997), who 'trimmed' the common support region of those treated individuals falling where the comparison group density, albeit strictly positive, is still considered to be too thin to produce reliable estimates.

In our empirical application, we use the publicly available STATA command that was developed by Leuven and Sianesi (2003) that performs various types of Mahalanobis metric and propensity score matching and allows us to impose common support in the ways that were described above as well as to test the resulting matching quality in terms of covariate balance in the matched groups.

3.2.3. *The multiple-treatment model*

Rosenbaum and Rubin's (1983) potential outcome approach for the case of a single treatment has recently been generalized to the case where a whole range of treatments are available by Imbens (2000) and Lechner (2001a). With MM1 and MM2 appropriately extended, all the required effects are identified. As with the single-treatment case, it is easy to show that a one-dimensional (generalized) propensity score can be derived, which ensures the balancing of the observables in the two groups being compared at a time.

3.2.4. *Some drawbacks to matching*

The most obvious criticism that may be directed to the matching approach is the fact that its identifying conditional independence assumption (MM1) is in general very strong. Despite the fact that, compared with OLS, matching is implemented in a more flexible way (in particular not imposing linearity or a homogeneous additive treatment effect), both matching and OLS estimates depend on this same crucial assumption of the selection on observables, and both are thus as good as the control variables X that they use (see also Smith and Todd (2004)). As mentioned above, the plausibility of such an assumption should always be discussed case by case, with account being taken of the informational richness of the available data set (\tilde{X}) in relation to a detailed understanding of the institutional set-up by which selection into the

treatment takes place (see Sianesi (2004) for an example of such a discussion in the context of training programmes).

Furthermore, the common support requirement that is implicit in MM2 may at times prove to be quite restrictive. In the case of social experiments, randomization generates a comparison group for each \tilde{X} in the population of the treated individuals, so that the average effect on the treated group can be estimated over the entire support of the treated group. By contrast, under the conditional independence assumption, matching generates a comparison group, but only for those \tilde{X} -values that satisfy MM2. In some cases, matching may not succeed in finding a non-treated observation with a similar propensity score for all the participants. If this assumption fails for some subgroup(s) of the participants, the estimated treatment effect must then be redefined as the mean treatment effect for those treated falling within the common support.

If the effect of treatment is homogeneous, at least within the treated group, no additional problem arises besides the loss of information. Note though that the setting is sufficiently general to include the heterogeneous case. If the effect of participation differs across treated individuals, restricting to the common support may actually change the parameter that is being estimated; in other words, it is possible that the estimated effect does not represent the mean treatment effect on the treated group. This is certainly a drawback of matching with respect to randomized experiments; when compared with standard parametric methods, though, it can be viewed as the price to pay for not resorting to the specification of a functional relationship that would allow us to extrapolate outside the common support. In fact, the absence of good overlap may in general cast doubt on the robustness of traditional methods relying on functional form (in the schooling context, see Heckman and Vytlacil (2000), Dearden *et al.* (2002) and Black and Smith (2004)). Lechner (2001b) has derived nonparametric bounds for the treatment effect to check the robustness of the results to the problem of a lack of common support.

3.3. Instrumental variable methods

The IV estimator seems a natural method to turn to in estimating returns—at least in the homogeneous returns model. The third source of bias in equation (8)—and the most difficult to avoid in the case of least squares and matching—arises from the correlation of observable schooling measures with the unobservables in the earnings regression. If an instrument can be found that is correlated with the true measure of schooling and uncorrelated with the unobservables in the outcome equation, then a consistent estimator of the returns is achievable in the homogeneous returns model but only in some special cases for the heterogeneous returns model. Even in the homogeneous returns model, though, finding a suitable instrument is not an easy task, since it must satisfy the criteria of being correlated with the choice of schooling while being legitimately excluded from the earnings equation.

To investigate the properties of the IV estimator more formally, consider the general heterogeneous model (6), which also allows for $b(X_i)$. Without loss of generality, this observably heterogeneous return $b(X_i)$ can be assumed to be linear in the X -variables, so that $b(X_i)S_{1i} = b_X X_i S_{1i}$, where b_X is the vector of the additional returns for individuals with characteristics X . Note again that, in this framework, b_i captures the individual idiosyncratic gain (or loss) and has population mean b_0 . The model can thus be written as

$$y_i = m_0(X_i) + b_X X_i S_{1i} + b_0 S_{1i} + e_i \quad \text{with } e_i = \alpha_i + \varepsilon_i + (b_i - b_0) S_{1i}. \quad (12)$$

Define an IV Z_i and assume that it satisfies the orthogonality conditions

$$E[\alpha_i | Z_i, X_i] = E[\varepsilon_i | X_i] = 0$$

(IV assumption 1, IV1) and

$$E[\varepsilon_i|Z_i, X_i] = E[\varepsilon_i|X_i] = 0$$

(IV assumption 2, IV2). With a valid instrument Z_i , we may envisage two ways of applying the IV method to estimate model (12).

- (a) *IV method (A)* uses the extended set of instruments Z_i and $Z_i X_i$ to instrument S_{1i} and $X_i S_{1i}$. It needs sufficient variation in the covariance of the interactions of X_i and S_{1i} and the interactions of X_i and Z_i . However, this approach does not fully exploit the mean independence assumptions IV1 and IV2.
- (b) *IV method (B)*, by contrast, recognizes that, under the conditional mean independence assumptions, application of the IV method is equivalent to replacing S_{1i} with its prediction in *both* its linear and its interactions terms. To see this, assume that

$$E[S_{1i}|Z_i, X_i] \text{ is a non-trivial function of } Z \text{ for any } X$$

(IV assumption 3, IV3). Taking the conditional expectation of model (12) under IV1–IV3 and noting that $E[X_i S_{1i}|Z_i, X_i] = X_i E[S_{1i}|Z_i, X_i]$ yields

$$E[y_i|Z_i, X_i] = m_0(X_i) + (b_X X_i + b_0) E[S_{1i}|Z_i, X_i] + E[(b_i - b_0) S_{1i}|Z_i, X_i]. \quad (13)$$

Note first that, in the absence of interactions $b(X_i)$, the two IV methods are identical. Secondly, irrespective of the method that is chosen, there is nothing in IV1–IV3 that makes the final term in equation (13) disappear. Since the error term e_i in model (12) contains the interaction between the endogenous schooling dummy variable and the unobserved individual return, neither way of applying the IV method would produce consistent estimates. In fact, even assuming that the instrument is uncorrelated also with the unobservable return component would not help further on its own. Two alternative paths can now be followed: considering some special cases based on further and stronger assumptions or redefining the parameter to be identified (specifically, as an LATE).

As to the first identifying strategy, one obvious possibility consists in assuming that returns are homogeneous, at least conditionally on X_i , i.e. that b_i is constant for all i and equal to its average value b_0 . Consequently, the problematic last term in equation (13) is zero by definition and, under IV1–IV3, IV estimation can produce a consistent estimator of $b(X_i) + b_0$. Note, however, how, in general, the IV estimator needs to deal with the specification of $m_0(X_i)$ and $b(X_i)$ and, just like least squares, is thus subject to the potential misspecification bias that the matching method avoids.

Special cases allowing for heterogeneous individual returns b_i and based on appropriate assumptions (e.g. homoscedastic returns) have been highlighted by Wooldridge (1997) for the one-factor ‘years-of-schooling’ specification. We now, however, focus on the general heterogeneous returns model with a single binary treatment (6).

3.3.1. Instrumental variables in the heterogeneous single-treatment model

As seen above, IV1–IV3 are not enough to ensure consistency in the general case of heterogeneous returns. IV3 requires that $E[S_1|Z, X] = P[S_1 = 1|Z, X]$ is a non-trivial function of Z for each X —in particular, it requires the instrument to take on at least two distinct values, say 0 and 1, which affect the schooling participation probability differently. Add now the additional property that, for the treated individuals, the instrument Z is not correlated with the individual-specific component of the return b_i (conditional on X). Formally

$$E[b_i|Z_i, X_i, S_{1i} = 1] = E[b_i|X_i, S_{1i} = 1]$$

(IV assumption 4, IV4). Under IV1–IV4, taking expectations, we obtain

$$E[y_i|Z_i, X_i] = m_0(X_i) + \{b(X_i) + E[b_i|X_i, S_{1i} = 1]\} P(S_{1i} = 1|Z_i, X_i),$$

from which we can recover the conditional effect of treatment on the treated group:

$$\begin{aligned} \hat{\beta}_{IV}(X) &\equiv \frac{E[y_i|X_i, Z_i = 1] - E[y_i|X_i, Z_i = 0]}{P[S_{1i} = 1|X_i, Z_i = 1] - P[S_{1i} = 1|X_i, Z_i = 0]} \\ &= b(X_i) + E[b_i|X_i, S_{1i} = 1] \\ &= E[y_i^1 - y_i^0|X_i, S_{1i} = 1] \equiv \hat{\beta}_{ATT}(X). \end{aligned} \tag{14}$$

IV4 is strong: although allowing for heterogeneous returns b_i , it requires schooling decisions to be unrelated to these individual gains. In particular, since IV3 requires the schooling participation probability to depend on Z , IV4 rules out that this probability depends on b_i as well.

Before turning to the issues that emerge when choices of schooling are allowed to depend on b_i , it is worth noting the issues of efficiency and of weak instruments. Efficiency concerns the imprecision that is induced in IV estimation when the instrument has a low correlation with the schooling variable. The weak instrument case is an extreme version of this where the sample correlation is very weak and the true correlation is near to zero. In this case, the IV method will tend to the biased OLS estimator even in very large samples (see Bound *et al.* (1995) and Staiger and Stock (1997)).

3.3.1.1. *The local average treatment effect.* In the general heterogeneous returns model with a single-treatment (6), even when individuals do partly base their choices of education on their individual-specific gain b_i , it is still possible to provide a potentially interesting interpretation of the IV estimator—although it does not estimate the average effect of treatment on the treated individuals or the average treatment effect parameters. The interpretation of IV in this model specification was precisely the motivation for the LATE of Imbens and Angrist (1994).

Suppose that there is a single discrete binary instrument $Z_i \in \{0, 1\}$ —e.g. a discrete change in some educational ruling that is positively correlated with the schooling level S_{1i} in the population. There will be four subgroups of individuals: those who do not take the education level under consideration whatever the value of the instrument (the ‘never takers’), those who always choose to acquire it (the ‘always takers’) and those who are induced by the instrument to change their behaviour, either in a perverse way (the ‘defiers’) or in line with the instrument (the ‘compliers’). This last group is of particular interest: it is made up of those individuals who would have education level $S_{1i} = 1$ after the rule change ($Z_i = 1$) but who would not have this level of schooling in the absence of the rule change ($Z_i = 0$). To be more precise, we define the events

$$\begin{aligned} D_{1i} &\equiv \{S_{1i}|Z_i = 1\}, \\ D_{0i} &\equiv \{S_{1i}|Z_i = 0\} \end{aligned}$$

and assume, in addition to the exclusion restrictions concerning the unobservables in the base state (IV1 and IV2) and to the non-zero causal effect of Z on S_{1i} (IV3—i.e. the instrument must actually change the behaviour of some individuals):

for all i , either $[D_{1i} \geq D_{0i}]$ or $[D_{1i} \leq D_{0i}]$ (note that, owing to IV3, strict inequality must hold for at least some i)

(LATE assumption 1, LATE1).

This ‘monotonicity’ assumption requires the instrument to have the same directional effect on all those whose behaviour it changes, *de facto* ruling out the possibility of either defiers or compliers. Assume in particular that $D_{1i} \geq D_{0i}$ (Z makes it more likely to take S_1 and there are no defiers); in this case, the standard IV estimator (14)

$$\frac{E[y_i|X_i, Z_i = 1] - E[y_i|X_i, Z_i = 0]}{P[S_{1i} = 1|X_i, Z_i = 1] - P[S_{1i} = 1|X_i, Z_i = 0]}$$

reduces to $b(X_i) + E[b_i|X_i, D_{1i} > D_{0i}] = E[y_i^1 - y_i^0|X_i, D_{1i} > D_{0i}]$. This provides a useful interpretation for IV: it estimates the average returns among those individuals (with characteristics X) who are induced to change behaviour because of a change in the instrument—the LATE.

More generally, the IV (two-stage least squares) estimator with regressors is a variance-weighted average of the LATEs conditional on the covariates. The IV estimator exploiting more than one instrument is an average of the various single-instrument LATE estimators with weights proportional to the effect of each instrument on the treatment dummy (see Angrist and Imbens (1995)).

LATEs avoid invoking the strong IV4. Indeed, as Angrist *et al.* (1996) noted, IV4 would amount to assuming that the return is the same for always takers and compliers—in other words, that it is the same for all the treated individuals, which comprise these two groups. However, if we are not willing to make this assumption, which would identify the ATT parameter as in equation (14), then the only causal effect to be identified by IVs is the LATE, i.e. the effect for compliers.

3.3.2. Some drawbacks to instrumental variables

The first requirement of IV estimation is the availability of a suitable and credible instrument. Although ingenious instruments have often been put forward (from selected parental background variables, to birth order, to smoking behaviour when young, to distance to college, etc.), they have all been subject to some criticism, since it is difficult to justify fully the untestable exclusion restriction that they must satisfy. Policy reforms have also been used as instruments. For example, researchers have compared the outcomes between two groups that have a similar distribution of abilities but who, from some exogenous reform, experience different schooling outcomes (for example, see Angrist and Krueger (1991, 1992), Butcher and Case (1994), Harmon and Walker (1995) and Meghir and Palme (2000)). As we have seen, in the homogeneous treatment effects model, this can be used to estimate the average treatment effect but, in the heterogeneous model where individuals act on their heterogeneous returns, it will estimate the average of returns among those who are induced to take more schooling by the reform—the LATE. The LATE discussion highlights the point that the IV estimate will typically vary depending on which instrument is used. Moreover, it could vary widely, when heterogeneity is important, according to the local average that it recovers, since the compliers could be a group with very high (or very low) returns.

In any case, the lesson that is to be learned from the discussion of IVs in the heterogeneous returns model is that the nature of the incidence of the instrument within the distribution of returns b_i is critical in understanding the estimated coefficient and may at times prove to be useful in bounding the returns in the population. A potentially promising approach in such a context (see, for example, Ichino and Winter-Ebmer (1999)) is to look for different instruments that are likely to affect different subgroups in the population, while having a theoretical framework to assess to which part of the returns distribution these complier groups belong. We provide some further discussion of this in relation to our application to the UK NCDS data in Section 4.2.

3.4. Control function methods

If individuals make educational choices on the basis of their unobserved characteristics, the error in the earnings equation will have a non-zero expectation (see equations (9) and (10)). In particular, if individuals who select into schooling have higher average unobserved ability and/or if individuals with higher unobserved idiosyncratic returns from schooling invest more in education, the residual in the earnings equation of high education individuals will have a positive mean. The basis of the control function approach is to recover the average treatment effect by controlling directly for the correlation of the error term in the outcome equation with the schooling variable (e_i in equation (9)). For this, an explicit model of the schooling selection process is required. More precisely, the control function method augments the earnings regression with an additional equation determining educational choice.

3.4.1. The single-treatment model

Suppose that, in the heterogeneous single-treatment model (6),

$$y_i = m_0(X_i) + \{b(X_i) + b_0\}S_{1i} + (b_i - b_0)S_{1i} + \alpha_i + \varepsilon_i,$$

assignment to schooling S_{1i} is determined according to the binary response model

$$S_{1i} = \mathbf{1}\{m_S(Z_i, X_i) + \nu_i \geq 0\} \quad \text{where } \nu_i \text{ is distributed independently of } Z \text{ and } X$$

(control function assumption 1, CF1). In addition to specifying this assignment rule, the control function approach requires that, conditionally on some function m_S , the unobservable heterogeneity in the outcome equation, α_i and b_i , is distributed independently of the schooling variable S_{1i} . One way of achieving this, in the single-treatment specification (6), is to assume that the unobserved productivity or ability term α_i and the unobserved individual residual return b_i relate to S_{1i} according to

$$\alpha_i - \alpha_0 = r_{\alpha\nu}\nu_i + \xi_{\alpha i} \quad \text{with } \nu_i \perp \xi_{\alpha i}$$

(control function assumption 2, CF2) and

$$b_i - b_0 = r_{b\nu}\nu_i + \xi_{b i} \quad \text{with } \nu_i \perp \xi_{b i}$$

(control function assumption 3, CF3). Note that generally—and as we do in our application below—joint normality of the unobservables in the assignment and outcome equations is assumed, from which CF2 and CF3 directly follow, with $r_{\alpha\nu} = \rho_{\alpha\nu}\sigma_\alpha$ and $r_{b\nu} = \rho_{b\nu}\sigma_b$.

Given CF1–CF3 we can write the conditional means of the unobservables as

$$\begin{aligned} E[(\alpha_i - \alpha_0) | Z_i, X_i, S_{1i} = 1] &= r_{\alpha\nu}\lambda_{1i}(X_i, Z_i), \\ E[(\alpha_i - \alpha_0) | Z_i, X_i, S_{1i} = 0] &= r_{\alpha\nu}\lambda_{0i}(X_i, Z_i), \\ E[(b_i - b_0) | Z_i, X_i, S_{1i} = 1] &= r_{b\nu}\lambda_{1i}(X_i, Z_i) \end{aligned} \tag{15}$$

where λ_{0i} and λ_{1i} are the conditional mean terms or ‘control functions’ that fully account for the dependence of the unobservable determinants of the outcome variable y on the schooling assignment. Consequently, the outcome model can be written as

$$\begin{aligned} y_i &= \alpha_0 + m_0(X_i) + \{b(X_i) + b_0\}S_{1i} + r_{\alpha\nu}(1 - S_{1i})\lambda_{0i} + (r_{\alpha\nu} + r_{b\nu})S_{1i}\lambda_{1i} \\ &\quad + \omega_i \quad \text{with } E[\omega_i | X_i, S_{1i}, (1 - S_{1i})\lambda_{0i}, S_{1i}\lambda_{1i}] = 0. \end{aligned} \tag{16}$$

If λ_{0i} and λ_{1i} were known, then the least squares estimation of the augmented log-earnings regression, which includes the additional terms $(1 - S_{1i})\lambda_{0i}$ and $S_{1i}\lambda_{1i}$, would produce a consistent estimator of the average treatment effect $b(X_i) + b_0$ and thus of $\beta_{ATE} = b_0 + E[b(X_i)]$. These

additional control function terms thus eliminate the bias that is induced by the endogeneity of schooling.

The control function terms depend on the unknown reduced form $m_S(\cdot)$ and the distribution of the unobservables. Under joint normality, the control functions take the form

$$\lambda_{0i} \equiv -\frac{\phi\{m_S(Z_i, X_i)\}}{1 - \Phi\{m_S(Z_i, X_i)\}},$$

$$\lambda_{1i} \equiv \frac{\phi\{m_S(Z_i, X_i)\}}{\Phi\{m_S(Z_i, X_i)\}}$$

and are the standard inverse Mills ratios from the normal selection model (Heckman, 1979). These can be consistently estimated from a first-stage binary response regression, which is analogous to the standard selection model. Once these terms have been included in the outcome equation (6) and implicitly subtracted from its error term $(b_i - b_0)S_{1i} + \alpha_i + \varepsilon_i$, the purged disturbance will be orthogonal to all the regressors in the new equation (see Heckman and Robb (1985)). For an early analysis of the heterogeneous one-factor years-of-schooling model, see, for example, Garen (1984). In general, an exclusion restriction on Z will allow semiparametric estimation of this model (see Powell (1994) for a review of semiparametric selection model estimation).

It is interesting to observe that, under the structure that is imposed on the model, the estimated r -coefficients are informative on the presence and direction of the selection process ($r_{\alpha\nu}$ for selection on unobserved ‘ability’ and $r_{b\nu}$ for selection on unobserved returns). Specifically, if an exclusion restriction can be found and the control function assumptions invoked, then the null hypothesis of no selection on the unobservables can be tested directly. In the framework above, this simply amounts to a joint test of the null hypothesis that $r_{\alpha\nu}$ and $r_{b\nu}$ are 0.

Not only does the model readily estimate the average treatment effect for a *random* individual *even* when individuals select into education on the basis of their unobserved individual gain from it (compare equation (15) with IV4), but also the distributional assumptions that are made allow us to recover the other parameters of interest:

$$\beta_{ATT} = b_0 + E[b(X_i)|S_{1i} = 1] + r_{b\nu}E[\lambda_{1i}|S_{1i} = 1],$$

$$\beta_{ATNT} = b_0 + E[b(X_i)|S_{1i} = 0] + r_{b\nu}E[\lambda_{0i}|S_{1i} = 0]$$

where $r_{b\nu}$ is identified from the difference in the coefficients on $S_{1i}\lambda_{1i}$ and on $(1 - S_{1i})\lambda_{0i}$. In the special case where b_i is constant for all i or where individuals do not select on the basis of their unobserved gain (b_i and ν_i are uncorrelated, so that $r_{b\nu} = 0$), the control function terms reduce to a single term $r_{\alpha\nu}\{(1 - S_{1i})\lambda_{0i} + S_{1i}\lambda_{1i}\}$.

In summary, although in general an exclusion restriction is required (see also Section 3.4.3 below), the structure that is imposed by the control function approach yields several gains compared with IVs. First, it allows us to recover the average treatment effect even when individuals select on the basis of unobserved heterogeneous returns; in such a context the IV method would by contrast be able to recover only an LATE for the specific and instrument-related subpopulation of compliers.

Secondly, whereas IVs only allow us to test the joint null hypothesis of no selection on either unobserved components of levels or gains, the control function structure allows us to test *separately* for the presence of selection on unobserved characteristics affecting the no-treatment outcome and for selection on unobserved heterogeneity in returns. These tests can be very informative in themselves. The former test can also give guidance on the reliability of the conditional

independence assumption on included covariates that underpins the matching specification. The latter test can also assist in the interpretation of IV estimates.

Finally, at times it may be important to allow for *observably* heterogeneous returns in addition to selection on unobservables. The available instruments may turn out to be too weak to predict all interactions properly. If X -heterogeneous returns are ignored, the IV method would again retrieve a local effect, whereas the control function would recover ATE, ATT and ATNT, and would do so in a considerably more efficient way—at the obvious price of being much less robust than IVs. These issues are further explored and discussed in our empirical application in Section 4.2.

3.4.2. *The multiple-treatment model*

The extension to the multiple-treatment case is reasonably straightforward. As in equation (7), write the exhaustive set of J treatments (schooling levels) under examination as $S_{1i}, S_{2i}, \dots, S_{Ji}$. Then extend the control function assumptions to obtain (where now a bar rather than a 0-subscript denotes means to avoid confusion)

$$\begin{aligned}
 E[(\alpha_i - \bar{\alpha})|Z_i, X_i, S_{ji} = 1] &= r_{\alpha\nu} \lambda_{ji}(X_i, Z_i) && \text{for } j = 0, 1, \dots, J, \\
 E[(b_{ji} - \bar{b}_j)|Z_i, X_i, S_{ji} = 1] &= r_{b_j\nu} \lambda_{ji}(X_i, Z_i) && \text{for } j = 1, \dots, J.
 \end{aligned}$$

The heterogeneous returns model specification is then given by

$$\begin{aligned}
 y_i &= \bar{\alpha} + m_0(X_i) + \sum_{j=1}^J \{b_j(X_i) + \bar{b}_j\} S_{ji} + \sum_{j=0}^J r_j S_{ji} \lambda_{ji} + \omega_i \\
 &\text{with } S_{0i} = 1 - \sum_{j=1}^J S_{ji}, r_j = r_{\alpha\nu} + r_{b_j\nu} \text{ for all } j \text{ (with } r_{b_0\nu} = 0)
 \end{aligned}$$

and

$$E[\omega_i | X_i, S_{1i}, \dots, S_{Ji}, S_{1i} \lambda_{1i}, \dots, S_{Ji} \lambda_{Ji}] = 0.$$

To avoid multicollinearity problems, the λ_{ji} -terms will need to have independent variation, suggesting that at least $J - 1$ excluded instruments will be required for identification. Typically, finding such a large set of ‘good’ excluded instruments is difficult. An alternative identification strategy is to link the λ_{ji} -terms together. For example, if the schooling outcomes follow an ordered sequence, then it may be that a single ordered probit model could be used to generate *all* the λ_{ji} -terms, but requiring only one instrument.

Within this multiple-treatment structure, all the treatment effects of interest can be obtained. For example, the generic average return to schooling level j compared with schooling level 0 (the return to which is normalized to 0) for those individuals with highest achieved schooling qualification k is

$$\begin{aligned}
 E[\beta_{ji} | X_i, S_{ki} = 1] &= ATE_{j0}(X_i) + r_{b_j\nu} E[\lambda_{ki} | S_{ki} = 1] \\
 &= b_j(X_i) + \bar{b}_j + r_{b_j\nu} E[\lambda_{ki} | S_{ki} = 1].
 \end{aligned}$$

3.4.3. *Some drawbacks to the control function approach*

In general, like the IV approach, the control function approach rests on an exclusion restriction. More precisely, although in a parametric specification identification can be achieved even if $X = Z$ through functional form restrictions, in practice the estimator is found to perform poorly in the absence of an exclusion restriction.

In contrast with the IV method, the control function approach also requires a full specification

of the assignment rule. These assumptions then allow the range of treatment effect parameters to be recovered even where there is heterogeneity in returns. The full relationship between the control function and IV approaches for general simultaneous models is reviewed in Blundell and Powell (2003). In the multiple-treatment model, a full set of assignment rules is required as well as the ability to construct a set of control functions—one for each treatment—that have independent variation.

3.5. *The relationship between ordinary least squares, matching, instrumental variables and control function methods*

This final subsection outlines the relationship between the estimators that we have considered. The emphasis of the matching approach is on the careful construction of a comparison group. The control function method aims at putting enough structure to model the selection decision completely, whereas IVs focus on the search for a source of independent variation affecting the choices of schooling of a section of the population.

To simplify the discussion, assume that

- (a) there is additive separability in unobservables, as in equation (3),
- (b) the issue of common support can be ignored (either by assuming that there is sufficient overlap in the distribution of X in the treated and non-treated subsamples or by assuming that all estimators condition on observations falling within the common support) and
- (c) there are no misspecification issues about the no-treatment outcome $m_0(X)$.

Of course, when these conditions fail, matching always dominates OLS. To consider further the relationship between standard OLS and matching, assume for the moment also that MM1 holds (i.e. no selection on unobservables). As shown in Section 3.1, under these conditions and in contrast with matching, standard OLS will still not recover ATT, although at times it might provide a close approximation, as shown by Angrist (1998). In particular, both matching and OLS produce weighted averages of the covariate-specific treatment effects $E[y^1 - y^0|X] \equiv b(X)$, but the ways that the two estimators weight these heterogeneous effects differ. Matching recovers ATT by weighting the X -heterogeneous effects according to the proportion of treated individuals at each value of X —i.e. proportionally to the propensity score at X , $P(S_1 = 1|X = x) \equiv p(x)$:

$$ATT \equiv E[Y^1 - Y^0|S_1 = 1] = \frac{\sum_x b(x) p(x) P(X = x)}{\sum_x p(x) P(X = x)},$$

By contrast, simple OLS weights the X -heterogeneous effects proportionally to the variance of treatment status at X —i.e. proportionally to $p(x)\{1 - p(x)\}$:

$$\beta_{OLS} = \frac{\sum_x b(x) p(x)\{1 - p(x)\} P(X = x)}{\sum_x p(x)\{1 - p(x)\} P(X = x)}.$$

In general, then, simple OLS will not recover ATT *even* under the conditional independence assumption and the conditions that are stated above. It will none-the-less provide a close approximation to ATT if there is no large heterogeneity in treatment effects by X or, alternatively, if all the values of the propensity score are smaller than 0.5 (hence p and $p(1 - p)$ are positively correlated).

For the remainder of the discussion, assume further that

- (d) the OLS, IV and control function estimators are properly specified, also in terms of $b(X_i)$ (a not-so-weak proviso, as we shall see in Section 4).

The four assumptions (a)–(d) rule out the two sources of bias due to observables B_1 and B_2 . Note first that OLS and matching now coincide. Secondly, once we have thus brought all estimators onto an equal footing, matching (equal to OLS), IVs and the control function would produce the same estimates in the absence of selection on unobservables. In what follows, we therefore look at a situation that is characterized by bias due to unobservables only (B_3).

To focus on the relative performance of matching compared with the IV and control function estimators when the basic conditions for the applicability of the latter are met, let us further assume that the exclusion restriction $E[\alpha_i | X_i, Z_i] = 0$ for the instrument used by the IV method as well as the decomposition that is required by the control function estimator (including postulated structure between the error terms and exclusion restriction) is verified.

In the presence of ability bias, arising from the correlation between α_i and S_{1i} , both the IV and the control function estimators should correctly recover the average effect of treatment on the treated group (IVs directly; the control function exploiting the assumed structure). The effect of treatment on the treated group recovered by matching would, however, be upwardly biased (assuming that more able individuals are more likely to choose $S_{1i} = 1$); the effect of treatment on the non-treated group would be similarly upwardly biased, and thus so would the average treatment effect.

When selection into schooling is driven by individuals' idiosyncratic gain, b_i , the control function estimator would directly recover the average treatment effect, whereas the IV method would pick out an instrument-related margin (the LATE), which could be much higher or much lower than the average effect for a random individual in the population. Provided that the individual-specific gain is unrelated to ability (α_i), both the matching and the control function estimators could recover an unbiased estimate of the average treatment effect on the treated group. However, in contrast with the control function estimate, the effect of treatment on the non-treated group—and thus the average treatment effect—that is obtained with matching would be upwardly biased (assuming that those with the higher gains select into education).

Finally, it is worth noting how we might use additional information on a credible exclusion restriction (conditional on the included conditioning variables \tilde{X}). Together with the control function assumptions—including separability in unobservables—this additional information can be used to 'test' the null hypothesis of no selection on unobservables. This relies on the truth of the exclusion restriction and would test for the significance of the additional control function terms conditionally on the exogenous (\tilde{X} -) variables. We pursue this approach in our empirical analysis in the next section.

4. Education and earnings in Britain: results from the National Child Development Survey

4.1. Introduction

The availability of birth cohort data in Britain presents an ideal basis for examining the issues that are involved in estimating the returns to education. Here, we use data from the NCDS, which keeps detailed longitudinal records on all the children who were born in a single week in March 1958. These data have been used extensively in the analysis of health, family and economic outcomes (see Fogelman (1983) and McCulloch and Joshi (2002), for example). The main surveys that we use were undertaken in 1965, 1969, 1974, 1981 and 1991. These include

Table 1. Summary statistics†

<i>Variable</i>	<i>Mean</i>	<i>Standard deviation</i>	<i>Variable</i>	<i>Mean</i>	<i>Standard deviation</i>
Real log(hourly wage) 1991	2.040	(0.433)	Mother's education missing	0.159	(0.366)
Qualifications			Father's age in 1974	43.17	(13.74)
O-levels or equivalent	0.821	(0.383)	Father's age missing	0.075	(0.263)
A-levels or equivalent	0.548	(0.498)	Mother's age in 1974	41.48	(10.86)
HE	0.283	(0.451)	Mother's age missing	0.049	(0.216)
White	0.969	(0.173)	Father's social class in 1974		
Mathematics ability at 7 years			Professional	0.044	(0.205)
5th quintile (highest)	0.212	(0.408)	Intermediate	0.145	(0.352)
4th quintile	0.190	(0.392)	Skilled non-manual	0.076	(0.265)
3rd quintile	0.185	(0.389)	Skilled manual	0.297	(0.457)
2nd quintile	0.158	(0.365)	Semi-skilled non-manual	0.010	(0.098)
1st quintile (lowest)	0.141	(0.348)	Semi-skilled manual	0.095	(0.293)
Reading ability at 7 years			Unskilled	0.029	(0.167)
5th quintile (highest)	0.165	(0.371)	Missing, or unemployed or no father	0.306	(0.461)
4th quintile	0.187	(0.390)	Mother employed in 1974	0.513	(0.500)
3rd quintile	0.188	(0.391)	Number of siblings	1.692	(1.789)
2nd quintile	0.179	(0.383)	Number of siblings missing	0.106	(0.308)
1st quintile (lowest)	0.166	(0.372)	Number of older siblings	0.821	(1.275)
Ability at 7 years missing	0.115	(0.319)	Father's interest in education		
Mathematics ability at 11 years			Expects too much	0.013	(0.114)
5th quintile (highest)	0.199	(0.399)	Very interested	0.252	(0.434)
4th quintile	0.179	(0.384)	Some interest	0.215	(0.411)
3rd quintile	0.157	(0.364)	Mother's interest in education		
2nd quintile	0.152	(0.359)	Expects too much	0.032	(0.175)
1st quintile (lowest)	0.122	(0.328)	Very interested	0.344	(0.475)
Reading ability at 11 years			Some interest	0.354	(0.478)
5th quintile (highest)	0.176	(0.381)	Bad finances in 1969 or 1974	0.159	(0.365)
4th quintile	0.176	(0.381)	Region in 1974		
3rd quintile	0.163	(0.369)	North West	0.100	(0.300)
2nd quintile	0.163	(0.369)	North	0.070	(0.256)
1st quintile (lowest)	0.132	(0.338)	East and West Riding	0.079	(0.270)
Ability at 11 years missing	0.191	(0.393)	North Midlands	0.072	(0.258)
Comprehensive school 1974	0.468	(0.499)	East	0.073	(0.261)
Secondary modern school 1974	0.162	(0.368)	London and South East	0.143	(0.350)
Grammar school 1974	0.099	(0.299)	South	0.057	(0.232)
Private school 1974	0.052	(0.222)	South West	0.061	(0.240)
Other school 1974	0.018	(0.134)	Midlands	0.088	(0.283)
Missing school information	0.201	(0.401)	Wales	0.054	(0.227)
Father's years of education	7.270	(4.827)	Scotland	0.096	(0.295)
Father's education missing	0.172	(0.377)	Other	0.107	(0.308)
Mother's years of education	7.342	(4.606)			

†Number of observations, 3639.

information on parents' education and social class, financial problems in the family in 1969 and 1974, mathematics and reading ability at ages 7 and 11 years, the type of school and detailed qualifications, teachers' assessments, and earnings, employment and training since leaving education. We use data from the waves up to the 1991 survey in which the individuals were aged 33 years. Summary statistics are presented in Tables 1 and 2.

We begin by looking at a simple single-treatment model and consider the returns to college *versus* no college, which in the UK context is the return from undertaking some form of HE. We subsequently consider a sequence of multiple treatments starting with no (or extremely low level) qualifications, O-levels or vocational equivalent, A-levels or vocational equivalent, or

Table 2. Sample sizes in the NCDS

<i>Sweep</i>	<i>Year</i>	<i>Age (years)</i>	<i>Sample</i>
0	1958	0	17419
1	1965	7	15496
2	1969	11	18285
3	1974	16	14761
4	1981	23	12538
5	1991	33	11363

some type of HE qualification (see Appendix A for details of our educational classification).

The outcome of interest is individual wages at age 33 years in 1991. To focus fully on the returns to education and to avoid issues that are associated with selection into employment, we restrict our attention to males. We do not expect substantive bias arising from measurement error in schooling owing to the relative accuracy of the NCDS education measure. Contrary to standard cross-sectional data sets relying on recall information, individuals in the NCDS are followed since birth and throughout their schooling period, with examination files being collected from schools and interviews being carried out very close to the dates of completion of education. A US finding that is relevant to our single-treatment analysis is by Kane *et al.* (1999). Using the National Longitudinal Study of the High School Class of 1972, they found that self-reported schooling measures are fairly accurate—in fact, more accurate than transcript measures—in discriminating between those who have not attended college and those who have completed their degree. Abstracting from additional concerns potentially arising from (non-classical) measurement error allows us in the following application to devote full attention to the issues that we have discussed at length in Section 3: selection, heterogeneous returns, misspecification and comparability of groups.

4.2. *Single-treatment models: higher education*

The estimated returns to undertaking some form of HE are shown in Table 3. In this model, the ‘non-treated’ are a heterogeneous group that is made up of those leaving school with no formal qualifications, those stopping at O-levels and those finishing with A-levels.

Some comments on the choice and interpretation of the control variables X may be useful at this stage. As described in Section 2, the X s need to be ‘attributes’ of the assignment rule and of the earnings process unaffected by the treatment itself. Suitable regressors are thus pretreatment variables, as well as all time invariant individual characteristics. All such variables that are thought to influence *both* the educational decision of interest *and* wage outcomes should ideally be included as regressors. IVs would not make good conditioning regressors. Finally, since our conditioning X -variables are measured before (or at the time of) the educational choice, the treatment effects that we estimate will include the effect of schooling on some subsequent X which would also affect measured outcomes (examples include on-the-job training, tenure, experience and type of occupation found). The treatment effect will thus consist of all channels through which education affects wages, both directly (e.g. through productivity) and indirectly (via some of the X s).

4.2.1. *Selection on observables: ordinary least squares and matching*

We begin by comparing the two methods that rely on the selection on observables assumptions—namely OLS and matching. We focus on the standard form of OLS, the linear and

Table 3. Returns to HE compared with less than HE: average treatment effect, ATE, average effect of treatment on the treated, ATT, and average effect of treatment on the non-treated, ATNT†

Method	% wage gain for the following effects:		
	ATT	ATE	ATNT
OLS			
(i) Basic specification	39.8 (37.1; 42.5)	39.8 (37.1; 42.5)	39.8 (37.1; 42.5)
(ii) Full specification	28.7 (25.7; 31.8)	28.7 (25.7; 31.8)	28.7 (25.7; 31.8)
(iii) Fully interacted	26.5 (23.0; 30.1)	30.8 (27.6; 34.1)	32.5 (28.9; 36.2)
Matching			
(iv) Basic specification	40.1 (37.5; 43.1)	40.1 (37.5; 42.8)	40.2 (37.5; 42.8)
(v) Full specification	26.8 (23.5; 31.1)	31.3 (28.7; 34.9)	33.1 (30.0; 36.7)
Control function (heterogeneous returns)			
(vi) Full specification	51.6 (27.9; 85.7)	37.4 (19.2; 61.9)	31.7 (14.0; 54.5)
(vii) Fully interacted	29.4 (9.8; 47.9)	22.0 (1.6; 36.6)	19.1 (-10.7; 38.2)
IVs			
(viii) Bad financial shock	117.1 (41.9; 192.3)		
(ix) Parental interest	60.6 (15.1; 106.1)		
(x) Presence of older siblings	5.2 (-70.8; 60.4)		

†Basic specification, ethnicity and region, implicitly gender and age; full specification, as basic, plus standard family background information, tests at 7 and 11 years and school variables; the family background variables are mother's and father's education, age, father's social class when the child was 16 years old, the mother's employment status when the child was 16 years old and the number of siblings the child had at 16 years of age; control function, parental interest as instrument, for row (vii) interacted with X in the first-step probit; sample size $N = 3639$, except for matching; ATE (3414), ATT (1019) and ATNT (2395); numbers in parentheses are the 95% confidence intervals based on White-corrected robust standard errors for all specifications except for (iv), (v) and (vii), for which the bootstrapped 95% bias-corrected percentile confidence intervals (500 repetitions) are reported.

common coefficient specification, of which nonparametric (or semiparametric) matching represents a flexible version. The choice of kernel-based matching over other types of matching estimators has been guided by indicators of the resulting balancing of X that are presented in summary form in Table 4. (The results were, in any case, very close.) Our comparison of the two methods also includes an assessment of their sensitivity to the richness of the conditioning data. Given their common identifying assumption, the nature of the available observables is crucial for the credibility of the estimates. In particular, we compare estimates that are based on the detailed information in the NCDS with those that were obtained from the standard pretreatment information in commonly available data sets. These are presented in Table 3.

Specification (i) in Table 3 gives the OLS estimate when we use only minimal controls (region and ethnicity). The corresponding matching estimate is shown in row (iv). We see that the estimated return to HE for men is around 40% for both estimators, with the matching point estimate very close to the estimate from OLS.

When we include a richer set of controls—ability measures at both 7 and 11 years of age, the type of school and standard family background variables (specifications (ii) and (v))—both these estimates fall at between 27% and 33%. In particular, the OLS coefficient—which is constrained to be homogeneous—shows an average wage gain of 28.7% from taking some form of HE.

Matching is more informative, showing that the higher educated enjoy an average gain of 26.8% from having taken HE (ATT), whereas the estimated return for those who stopped (at any stage) before HE would have been 33.1% (ATNT).

Table 4. Covariate balancing indicators before and after matching (best specification)

Treatment	N_1 , before	Comparison	N_0 , before	Probit pseudo- R^2 , before (1)†	Probit pseudo- R^2 , after (2)‡	$P > \chi^2$, after (3)§	Median bias, before (4)§§	Median bias, after (5)§§	% lost to common support, after (6)*
HE	1030	No HE	2609	0.209	0.006	0.9963	9.1	1.4	0.0
No HE	2609	HE	1030	0.209	0.037	0.0000	9.1	3.4	0.8
None	651	O-level	993	0.150	0.005	1.0000	9.0	1.1	0.0
		A-level	965	0.248	0.012	0.9985	13.4	1.7	3.0
O-level	993	HE	1030	0.512	0.091	0.0000	15.1	6.0	5.5
		None	651	0.150	0.016	0.9491	9.0	2.7	1.2
		A-level	965	0.045	0.002	1.0000	6.5	0.6	1.3
A-level	965	HE	1030	0.227	0.019	0.4570	11.4	2.9	0.3
		None	651	0.248	0.041	0.0005	13.4	4.1	7.7
		O-level	993	0.045	0.002	1.0000	6.5	0.7	1.1
HE	1030	HE	1030	0.127	0.008	0.9999	7.6	1.5	0.5
		None	651	0.512	0.162	0.0000	15.1	10.3	20.2
		O-level	993	0.227	0.022	0.1906	11.4	2.9	5.9
		A-level	965	0.127	0.005	1.0000	7.6	1.4	0.5

†Pseudo- R^2 from probit estimation of the conditional treatment probability, giving an indication of how well the 52 regressors X explain the relevant educational choice.

‡Pseudo- R^2 from a probit of D on X on the *matched* samples, to be compared with column (1).

§ P -value of the likelihood ratio test after matching, testing the hypothesis that the regressors are jointly insignificant, i.e. well balanced in the two matched groups.

§§Median absolute standardized bias before and after matching, with median taken over all the 52 regressors. Following Rosenbaum and Rubin (1985), for a given covariate X , the standardized difference *before* matching is the difference of the sample means in the full treated and non-treated subsamples as a percentage of the square root of the average of the sample variances in the full treated and non-treated groups. The standardized difference *after* matching is the difference of the sample means in the matched treated (i.e. falling within the common support) and matched non-treated subsamples as a percentage of the square root of the average of the sample variances in the full treated and non-treated groups:

$$B_{\text{before}}(X) \equiv 100 \frac{\bar{X}_1 - \bar{X}_0}{\sqrt{[V_1(X) + V_0(X)]/2}}$$

$$B_{\text{after}}(X) \equiv 100 \frac{\bar{X}_{1M} - \bar{X}_{0M}}{\sqrt{[V_1(X) + V_0(X)]/2}}$$

Note that the standardization allows comparisons between variables X and, for a given variable X , comparisons before and after matching.

*Share of the treated group falling outside the common support, imposed at the boundaries and, in the multiple-treatment case, across all transitions.

As seen in Section 3.1, if there are heterogeneous returns to HE, standard OLS regression would, in general, produce biased estimates of ATT. To check this issue, in specification (iii) we run a regression that models the (observably) heterogeneous returns $b(X_i)$ in a flexible way, namely it allows *all* interactions between the X 's and the treatment indicator S_1 . These interactions $X S_1$ —particularly in terms of later ability, family background and region—are significant (overall $F = 1.80$; $p = 0.0019$), and allowing for them makes the OLS estimates of ATT, ATNT and ATE almost identical to the matching estimates (compare specification (iii) with specification (v)).

This first set of results highlights several issues. At least in our application, the standard pre-education information that is available in common data sets would not have been enough to identify gains in a reliable way; in our case, generally unobserved ability and family background variables would have led to an upward bias of around 48%.

Secondly, allowing for an (observably) heterogeneous gain from HE via matching or fully interacted OLS can, in principle, provide additional information on the average gains for the subgroups of treated and non-treated individuals. The statistical significance of the interaction terms provides evidence of the presence of heterogeneous returns $b(X_i)$. Furthermore, such heterogeneity seems to be sizable; both the interacted OLS and matching estimates of ATT are significantly different from the corresponding ATNT estimates. (The bootstrapped 95% bias-corrected percentile confidence interval for the -6.3 difference in matching estimates is $[-9.9; -2.6]$ and that for the -6.0 difference in interacted OLS estimates is $[-10.4; -2.1]$.) The results appear to imply that, if those who did not continue to HE had instead undertaken it, they would have enjoyed a substantially higher benefit than the group who effectively went on to HE.

Before taking these results on the average effect on the non-treated group at face value, important *caveats* need, however, to be considered, all of which point to a likely *upward* bias of this estimate. As seen in Section 3.2.1, identification of ATNT requires more restrictive assumptions—in particular, no selection based on unobserved returns. If this assumption is violated, and assuming that those with the higher gains select into HE, the matching estimate of ATNT would be upwardly biased. However, leaving selection on unobservables aside until the next subsection, it can easily be checked that matching does not perform well in balancing the X s. Table 4, column (3), shows that, in sharp contrast with the case of ATT, for ATNT a test of the hypothesis that the X s are well balanced in the two matched groups is rejected at any significance level (experimenting with a more flexible specification of the propensity score did not improve balancing; nor did it change the point estimate). Note that the non-treated group—a much larger group than the HE group of potential comparisons—also contains all those individuals who dropped out at 16 years of age without any qualifications. To calculate ATNT, these drop-outs need all to be matched to the most ‘similar’ HE individuals. By contrast, when estimating ATT for those with HE, the matching algorithm was free not to use those no-qualifications individuals who were not the best matches for the HE individuals; indeed, in the one-to-one version of the estimator, individuals with A-levels make up 53.6% of the matched comparisons, individuals with O-levels 37% and individuals with no qualifications only 9.4%. Considerable initial differences between these two groups would make it difficult to obtain reasonably good matches. In fact, one can easily verify how test scores remain badly unbalanced—in particular, there are far fewer low scoring matched HE individuals than there are no-HE treated individuals. Matching thus did not succeed in choosing an HE subgroup that looked as ‘low performing’ as the full no-HE group; hence, we know that ATNT from matching is upwardly biased just from considering the observables. Incidentally, since the average treatment effect is an average of the ATT- and ATNT-parameters, it will be affected by a poorly estimated ATNT.

Fully interacted OLS produced very similar point estimates as well as confidence intervals to those produced by matching. However, a flexible but parametric method such as our fully interacted OLS would have hidden from the analyst the fact that observationally different individuals were *de facto* being compared on the basis of extrapolations purely based on the functional form imposed.

This discussion draws attention to how matching estimators can, by contrast, appropriately highlight the problem of common support and thus the actual comparability of groups of individuals (see also Heckman *et al.* (1999)). Both matching and OLS deal with observables only; matching, however, also offers simple and effective ways of assessing *ex post* the quality of a matched comparison group in terms of the observables of interest. Nonparametric (or semi-parametric) methods such as matching thus force the researcher to compare only comparable individuals. If, however, the treated and non-treated groups are too different in terms of the observables, the researcher needs to accept the fact that there simply is not enough information

in the available data to achieve sufficiently close—and thus reliable—matches. This, in fact, turned out to be so for our ATNT estimate.

As for ATT, note that the matching and simple OLS estimates are very close (and in fact not significantly different). As we discussed in Section 3.5, in a given application we would expect little bias for ATT from simple OLS *vis-à-vis* matching if there is

- (a) no common support problem,
- (b) little heterogeneity in treatment effects according to X or, alternatively, all the propensity scores are ‘small’ (in particular, less than 0.5, which would make the weighting scheme of OLS proportional to the scheme for the matching estimator of ATT—see Angrist (1998)) and
- (c) no serious misspecification in the no-treatment outcome.

In fact, in our data, the common support restriction is not binding for ATT (see Table 4, column (6)), and only around 10% of the propensity scores in our sample are larger than 0.5. Hence if our specification of $m_0(X)$ is reasonably correct, we would expect matching and simple OLS to produce comparable estimates of ATT. Note, however, that matching dominates simple OLS *a priori*. Matching can quickly reveal the extent to which the treated and non-treated groups overlap in terms of pretreatment variables, it offers easy diagnostic tools to assess the balancing achieved and it relieves the researcher from the choice of the specification of $m_0(X)$. For the ATT in our data, it just turned out that these issues did not pose any serious problem; *a priori*, however, we could not have known how informative the data were.

4.2.2. Selection on unobservables: control function and instrumental variables

Both the OLS and the matching methods rely on the assumption of selection on observables. However rich our data set may seem, this is a strong assumption. IVs and control function approaches attempt to control for selection on unobservables by exploiting some ‘exogenous’ variation in schooling by way of an excluded instrument. The choice of an appropriate instrument Z , like the choice of the appropriate conditioning set X for matching or OLS, boils down to an untestable prior judgment. In fact, although there might be widespread consensus in including test score variables as ability measures among the X s or in viewing an exogenous change in some educational rule or qualification level for one group but not another as an appropriate instrument, ultimately the validity of the IV is untestable.

Our data contain some potential excluded variables that may determine the assignment to schooling but, conditional on the X s, may be excluded from the earnings equation, in particular birth order, father’s, mother’s and parents’ interest in the child’s education at age 7 years and adverse financial shocks hitting the child’s family at ages 11 and 16 years. All these ‘instruments’ are highly significant determinants of the choice to undertake HE (conditional on the full set of controls X), with individual F -values ranging from 8.3 to 18. However, we could of course still argue that, in addition to educational attainment, these ‘instruments’ could affect other individual traits (e.g. motivation or self-esteem) that could in turn affect earnings. Note that in our X -set we include ability (measured at 7 and 11 years of age) and standard family background controls; thus we require the instrument to be excluded from potential earnings for given ability, early school performance, family background and type of school.

Interestingly the control function, using these excluded instruments, cannot reject the hypothesis of no selection bias on unobserved ability (ρ_{ov}) conditional on the inclusion of the test score variables. A second informative result concerns the possibility of individuals selecting into HE on the basis of their idiosyncratic gains. In the richest specification, in addition to selection on unobserved ability, we allow for selection on observable heterogeneity in returns via the interactions

XS_1 as well as on unobservable heterogeneity in returns via a second control function term (an illustration with parental interest as the instrument is presented in row (vii) in Table 3). F -tests on the interaction terms indicate that there is indeed heterogeneity in returns according to X ($F = 1.47$; $p = 0.016$), whereas ρ_{bv} is not statistically significant. In contrast, when we do not control for (potentially) observable heterogeneity in returns (e.g. row (vi)), we can reject the hypothesis of no selection on unobserved returns. When ρ_{bv} is significant, it is thus picking up some misspecification in terms of the X s; once we control for heterogeneous returns in terms of our (rich) observables, from the control function specification there no longer appears to be any remaining selection on unobserved returns.

The control function estimates in Table 3 yield a point estimate of ATNT that is substantially lower than that of ATT. We have already argued that the ATNT that is estimated by matching and interacted OLS should not be regarded as a reliable measure of what non-graduates would have gained from taking HE. The structure that is imposed by the control function seems, by contrast, to yield results that are more consistent with individual maximizing behaviour, albeit much less precisely estimated.

If we have additional information in the form of an exclusion restriction we can utilize the IV estimator to construct an alternative check on the conditional independence assumption. Under the further assumption of no selection on unobserved individual gains, IVs on the fully interacted model should recover the average effect of treatment on the treated group.

To see how this works consider a factor M —say parental education—which is related to unobserved productivity and to the returns to HE, and which also enters the HE participation decision. In other words, M affects HE participation, the outcome y directly (conditional on X) and the returns from HE in terms of y .

- (a) When M is *unobserved*, provided that the instrument is uncorrelated with M given X (i.e. it satisfies exclusion restriction IV1 with respect to the unobservable M), the IV approach identifies an instrument-determined local effect, the LATE. If, however, the instrument is correlated with M , violating IV1, even the LATE would not be consistently estimated.
- (b) When M is *observed and we condition on it linearly in both the participation and the outcome equations*, since we do not control for the interaction, the IV approach is inconsistent: $M \cdot S_1$ is omitted from the outcome equation, giving rise to an omitted endogenous variable bias.
- (c) When M is *observed, we control for it linearly in the participation equation and interacted with S_1 in the outcome equation*, since there is now an additional $M \cdot S_1$ endogenous term in the outcome equation, we would need the additional instrument $M \cdot Z$. The potential problem in this case is that our instruments Z and $M \cdot Z$ may not predict the interactions S_1 and $M \cdot S_1$ very well, resulting in a loss of precision. The alternative of exploiting the schooling prediction both linearly and interacted with M may add some efficiency but would still require the instrument Z to provide sufficient variation in \hat{S}_1 and $M \cdot \hat{S}_1$ to allow b_0 and b_m to be independently identified. Either IV method would place strong demands on the instrument, particularly when there are many interaction terms.

Our previous findings that the return to HE does depend on the X s and that these X s also impact on the schooling decision would require our IV estimation to control for these endogenous interaction terms as well, i.e. according to case (iii) outlined above. By allowing observable heterogeneity in returns in a fully interacted model with all the XS_1 -terms, however, the estimates (which are not shown) are extremely imprecise. This severe lack of precision points to the fact that, although this IV estimation requires us to instrument every one of the endogenous

$X\hat{S}_i$ -terms, our corresponding instruments do not have enough power to predict all the interactions well, resulting in a poor performance of our interacted IV model. In fact, when we try to allow for X -heterogeneous returns, it is clear that our interacted instruments simply do not have enough power to identify our model (from the first-stage regressions and in particular from their ‘partial R^2 ’ Shea (1997) measure of instrument relevance that takes intercorrelations among instruments into account). We also experimented with controlling for the interactions by using IV method B, i.e. fully exploiting the conditional mean independence assumption and using the schooling prediction to replace both schooling and its interaction terms in the outcome equation (which is not reported). Although this does shrink the confidence interval, there still remains insufficient variation in \hat{S}_i and $X_i\hat{S}_i$ to recover a precise and statistically significant estimate of the average effect on the treated individuals.

However, from our control function results, we also know that, if we do not allow for heterogeneity in returns in terms of our X s, there will be selection on uncontrolled-for returns. As discussed in Section 3.3.1, in such a context of heterogeneous and acted-on returns, our simple IV estimates should be interpreted as estimates of LATEs: the average return to HE for those who go on to HE *because* of the change in the instrument. In this regard, we present results based on three different instruments that are likely to affect distinct subgroups of the population. Card (1999) has provided us with the theoretical framework for gauging where in the returns distribution these groups are likely to belong.

In Card’s model of endogenous schooling, individuals invest in education until the marginal return to schooling is equal to their marginal cost and where both marginal returns and costs are allowed to depend on schooling and to be heterogeneous. The causal effect of education on individual earnings (defined, for each individual, as the marginal return to schooling at that individual’s optimal schooling choice) is given in this model by

$$\beta_i = \theta b_i + (1 - \theta)r_i \quad (17)$$

where b_i captures differences in individuals’ returns due to ability (comparative advantage), r_i reflects differences in the opportunity costs that individuals face (e.g. their taste for schooling, individual discount rates and liquidity constraints) and θ is a constant in $[0, 1]$.

Assume for simplicity that there are only two values for each heterogeneity parameter, $b_H > b_L$ and $r_H > r_L$; the population is thus made up of the four types of individuals {LH, HH, LL, HL}. Then, from equation (17), we have the following (imperfect) ordering of the four returns:

$$\beta_{HH} > \{\beta_{LH}, \beta_{HL}\} > \beta_{LL}.$$

We consider three alternative instruments, observed variables that affect schooling choices but are uncorrelated with the ability factors in the earnings function and in the individual marginal return (thus effectively having to affect only the individual marginal cost r_i). Along the lines of Ichino and Winter-Ebmer (1999), we ask who the switchers for each instrument are. All the IV estimates in Table 3 are highly imprecise. But we still might ask what kind of interpretation would lead to the ranking of estimated local average returns that are indicated by the point estimates.

4.2.2.1. Adverse financial shock experienced by the family when the child was 11 or 16 years old. Individuals in the rich dynasties LL and HL suffer limited liquidity constraints; they always go on to HE irrespective of the shock. Individuals in the poor dynasty LH are subject to liquidity constraints and in addition are of low ability; they never take HE. By contrast, suffering

a bad financial shock affects the schooling for individuals in group HH, the high ability but liquidity-constrained individuals who choose to undertake HE only in the absence of the shock. We therefore expect our IV estimate based on financial shock to reflect the highest returns in the population, β_{HH} (117%—row (viii) in Table 3).

4.2.2.2. *Parental interest in the child's education at age 7 years (as perceived by the child's teacher)*. Individuals in the rich HL and LL dynasties would undertake HE independently of parental interest; similarly, the HH individuals, though poor, are of high ability and would go on to HE quite irrespective of their parents' interest in their education. By contrast, having parents who are very interested in one's education causes an increase in schooling for the LH individuals: they are of low ability and liquidity constrained and would never continue to HE unless they were pushed by their eager parents, who attach high value to education. The IV estimate that is based on parental interest would thus reflect some intermediate returns in the population, β_{LH} (60.6%—row (ix) in Table 3).

4.2.2.3. *Older siblings*. Individuals in the HL and HH dynasties have high ability and would continue into HE independently of the presence of older siblings. Those in the LH group are of low ability and severely liquidity constrained, and would thus never continue into HE. By contrast, the rich but low ability individuals LL would be those who are pushed by their rich family to obtain a degree only if they are the only (or the first-born) children in their family. Under this interpretation, the IV estimate based on the presence of older siblings would thus reflect the lowest returns in the population, β_{LL} (an insignificant 5.2%—row (x) in Table 3).

By considering different instruments that are believed to affect subgroups in given ranges of returns, we can thus gauge some (albeit quite imprecisely estimated) information concerning the extent of variability in HE returns in the population; we cannot, by contrast, retrieve information on average treatment effects owing to the lack of power of our instruments in predicting all the heterogeneous returns $X S_1$. By contrast, our control function with interactions model allows us to settle on the intermediate case, where all the $X S_1$ -interactions are included in the outcome equation and the $X Z$ -terms are exploited in the first-step probit, from which, however, still *only two* predictions (λ_1 and λ_0) need to be computed. All this is only possible by making stronger assumptions—in particular, we require additive heterogeneity in the unobservables. By placing more structure on the problem than does the IV approach, the control function method thus allows us to recover the ATE, ATT and ATNT parameters directly.

4.2.3. *Lessons and results from the single-treatment estimates*

In the NCDS, there appears to be some evidence suggesting that there are enough variables to be able to control *directly* for selection on unobservables—both unobservable individual traits and unobservable returns. In other words, we could not find any strong evidence that OLS and matching with the available set of X s are subject to selection bias; nor do individuals seem to select into HE on the basis of returns that are still unobserved by the econometrician. Connected to the latter point, we have found some evidence that interactions matter. More precisely, there is significant heterogeneity in returns to HE (especially in terms of parental education and region). In practice, though, the way that the heterogeneous returns are weighted by matching and by simple OLS turned out to be proportional in this application, resulting in simple OLS to recover the average effect on a treated individual fortuitously (see Section 3.5).

Both matching and fully interacted OLS resulted in an estimate of ATNT that is significantly higher than the estimate of ATT. However, we argued that such methods were most likely to yield upward-biased estimates of ATNT in this case.

4.3. Multiple-treatment models

We now turn to a more disaggregated analysis that focuses on the sequential nature of educational qualifications. We separate the qualifications variable into that for individuals who dropped out of school with no qualifications, that for those who stopped education after completing O-levels or equivalent, that for those who stopped after completing A-levels or equivalent and that for those who completed O-levels, A-levels and HE.

Since now we have four treatments, IV estimation would require at least three credible instruments. As to the control function approach, in the first stage we could exploit the sequential nature of the treatments and estimate an ordered probit model for the various levels of education based on one instrument only. This would, however, unduly rely on the (arbitrarily) imposed structure of the problem, since the model would be purely identified from the postulated treatment choice model. Instead in this section, we follow the conclusion of the previous discussion of the single-treatment model and assume that there are enough variables to be able to control directly for selection through matching.

Our approach involves estimating the incremental return to each of the three qualifications by actual qualification. For those with no qualifications, we estimate the returns that they would have obtained if they had undertaken each of the three qualifications (ATNT). For those with O-level qualifications, we estimate the return that they obtained for taking that qualification (ATT) and the returns that they would have obtained if they had progressed to A-levels or HE (ATNT). For those with A-levels, we estimate the returns that they obtained for undertaking O- and A-level qualifications (ATT) and the returns that they would have obtained if they had progressed to HE (ATNT). For those with HE, all estimates are ATNTs.

Our matching estimator adapts the estimation of the propensity score to the case of multiple sequential treatments (see Sianesi (2002) for more details). Outcomes across each of the four groups, matched on the appropriate propensity score for the particular transition in question, are then compared. Again we let the choice between various types of matching estimators be guided by how well they balanced the observed characteristics (see Table 4; in most cases Epanechnikov kernel matching performed best, though dominated for some comparisons by Mahalanobis metric matching).

The multiple-treatment results are shown in Table 5, where the estimates are those obtained from the 'best' specification (i.e. the specification resulting in the 'best' balancing of our X s in the matched subsamples) and after imposing the common support. Indeed, in estimating the effects and in calculating the probabilities for the average treatment effects, common support was imposed also in terms of only including individuals who are matched for every possible transition (so that we can make comparisons across the same sets of individuals). On the basis of the balancing of the observables within the matched samples (which is summarized in Table 4) we have highlighted the most (and the least) reliable results in Table 5.

As was the case in the single-treatment model, our first result is that controlling for ability and family background is important and reduces the return to education at all levels (compare the two OLS specifications; for brevity, only the results for the full set of controls X are presented for matching). Nevertheless, the findings show significant overall returns to educational qualifications at each stage of the educational process, even after correcting for detailed

Table 5. Incremental treatment effects by highest qualification achieved: matching and OLS estimates†

Educational group	% wage gain for the following comparisons:						N
	O-level versus none	A-level versus O-level	A-level versus none	HE versus A-level	HE versus O-level	HE versus none	
None	13.2 (9.1; 17.3)	5.5 (0.1; 10.1)	18.7 (13.6; 23.2)	24.8 (17.7; 31.6)	30.3 (23.2; 36.3)	43.5 (36.8; 49.7)	624
O-level	17.8 (12.9; 22.1)	5.9 (2.3; 9.9)	23.7 (19.1; 29.4)	24.6 (20.5; 29.3)	30.5 (26.6; 34.4)	48.2 (43.4; 53.3)	963
A-level	18.1 (13.2; 22.6)	5.7 (2.0; 9.8)	23.8 (18.5; 28.7)	25.6 (21.7; 30.2)	31.3 (27.5; 35.5)	49.4 (43.5; 54.0)	911
HE	21.6 (14.1; 29.6)	8.0 (3.9; 12.6)	29.6 (22.0; 37.5)	21.7 (17.4; 25.6)	29.7 (25.8; 33.7)	51.3 (43.8; 58.7)	871
Any: ATE	18.0 (13.3; 22.4)	6.3 (2.9; 10.1)	24.2 (19.7; 28.7)	24.2 (20.6; 28.2)	30.5 (27.1; 34.2)	48.4 (43.2; 52.7)	3251
OLS	14.8 (11.2; 18.4)	6.4 (3.1; 9.7)	21.2 (17.3; 25.1)	23.5 (20.0; 27.1)	29.9 (26.5; 33.4)	44.7 (40.1; 48.9)	3639
OLS basic	21.1 (17.4; 24.7)	9.0 (5.6; 12.4)	30.0 (26.2; 33.8)	28.9 (25.6; 32.3)	37.9 (34.7; 41.1)	59.0 (55.3; 62.6)	3639

†Controlling for ethnicity, region, standard family background information, tests at 7 and at 11 years and school variables; OLS basic, controlling for ethnicity and region only. Matching estimates, based on the 'best' specification, always imposing common support at the boundaries; common support is also imposed throughout all transitions; see Table 4 for the share of the treated group falling outside the common support in each comparison; numbers in parentheses are 95% bias-corrected percentile confidence intervals obtained by bootstrapping for the matching estimates (500 repetitions) and, for OLS, 95% confidence intervals based on robust standard errors; numbers in bold indicate the most reliable effects (based on balancing of the Xs between the groups—see Table 4) and those in italics are the least reliable effects.

background variables and differences in ability, as well as allowing for (observed) heterogeneity in the education response parameters.

OLS estimates and the average treatment effect that is obtained via matching are rather close, although OLS estimates are typically on the lower side. Matching shows an average wage return of 18% from obtaining O-levels compared with leaving school with no qualifications, a further 6% return from completing A-levels and a further 24% wage premium for then achieving HE. Compared with leaving school at 16 years of age without qualifications, then, the average return to O-levels is 18% and to A-levels is 24%, which doubles to 48% for HE. On an annualized basis, given that most individuals complete HE at 21 years of age, the average return is therefore 9.5%. If, by contrast, we just make the comparison with those who left at 16 years of age regardless of qualifications, the average total return would fall to 33.3% (the result is not shown), or 6.6% per annum. Similarly, the return per annum for A-levels (the achievement of which generally takes place at 18 years of age) compared with leaving at age 16 years without qualifications would be 12%, but only 3% if compared with stopping with O-levels at 16 years of age. A more adequate yearly measure is probably the measure that is obtained when the base-line comparison is with leaving school at 16 years of age irrespective of qualifications, which points to a yearly return to completing A-levels of 5.6%.

The set of results just discussed highlights the potential shortcomings of the one-factor years-of-schooling model (5) when applied to the UK educational system, in which individuals with the *same* number of years of schooling have quite different educational outcomes (no

Table 6. Difference in returns for the no-qualifications group†

<i>Comparison</i>	<i>Difference</i>	<i>95% confidence interval</i>
Returns to O-level <i>versus</i> none		
(a) For none <i>versus</i> for O-level	-4.5‡	[-9.3; -0.6]
(b) For none <i>versus</i> for A-level	-4.9‡	[-10.0; -0.2]
(c) For none <i>versus</i> for HE	-8.4‡	[-16.6; -0.2]
(d) For none <i>versus</i> for all (ATE)	-4.7§	[-9.2; -1.4]
Returns to A-level <i>versus</i> none		
(a) For none <i>versus</i> for O-level	-5.0‡	[-9.9; -1.2]
(b) For none <i>versus</i> for A-level	-5.1‡	[-11.4; -0.4]
(c) For none <i>versus</i> for HE	-10.9§	[-20.8; -3.8]
(d) For none <i>versus</i> for all (ATE)	-5.5§	[-10.3; -1.8]
Returns to HE <i>versus</i> none		
(a) For none <i>versus</i> for O-level	-4.7‡	[-11.0; -0.3]
(b) For none <i>versus</i> for A-level	-5.9§§	[-13.0; 0.1]
(c) For none <i>versus</i> for HE	-7.8§§	[-17.2; 1.1]
(d) For none <i>versus</i> for all (ATE)	-4.9§§	[-10.4; 0.1]

†Difference in returns to O-levels, A-levels and HE *versus* none for the no-qualifications group compared with (a) those who stopped at O-levels, (b) those who stopped at A-levels, (c) those who obtained HE and (d) the average treatment effect (i.e. across all four educational groups); the 95% bias-corrected percentile confidence intervals were obtained by bootstrapping.

‡Significant at the 5% level.

§Significant at the 1% level.

§§Significant at the 10% level.

qualifications or O-levels) and in which, irrespective of the comparison state that is chosen, imposing equality of yearly returns across educational stages proves too restrictive.

From the disaggregated results in Table 5, it appears that for the O-levels and A-levels groups, (observable) heterogeneity in impacts does not seem to be a particularly important feature of the data, so the point estimates for the two groups are extremely close and basically coincide with the corresponding estimates of the average treatment effects. By contrast, noteworthy new information arises for the (base-line) group of individuals who left school without any qualifications. For this group, the average returns to each educational investment (O-levels, A-levels and HE) compared with none would have been consistently the lowest among all the educational groups (Table 6), which might contribute to explaining their decision not to take any formal qualifications. None-the-less, if we focus on the returns to O-levels by educational group (first column), our disaggregated analysis shows that at that stage, even if those who do acquire some qualification at 16 years of age have the greatest returns from this initial investment, those who drop out at 16 years without any qualifications would still have had a hefty average pay-off of over 13% from obtaining O-levels or the equivalent before leaving education. Note that this result is obtained after controlling for detailed ability and family background. Furthermore, work by Harmon and Walker based on a natural experiment is consistent with this finding. Harmon and Walker (1995), exploiting changes in the minimum school-leaving age in the UK, found a return to schooling of 15–16%, whereas in later work (Chevalier *et al.*, 2002) they showed that, for men born ± 5 years around our NCDS cohort, the effect of the reform was solely in terms of a movement from no to low qualifications.

Individuals undertaking some form of HE are the second educational group to show considerable heterogeneity in returns, which thus visibly differ from the average treatment effects. In

particular, at 51.3%, they enjoy the highest overall return to HE (*vis-à-vis* no qualifications). The disaggregated results show that this higher average effect of HE for the HE-treated group actually stems from HE individuals enjoying a higher return from their initial O-level investment (21.6% compared with 13–18% for the other groups). In fact, their incremental return from A-levels to HE is lower than for those who did not undertake HE.

The results arising from comparing HE graduates with individuals without any qualifications (average returns to HE compared with no qualifications for the HE group, as well as returns derived from such an estimate) must, however, be viewed with great care. As summarized in Table 4, the HE and no-qualifications groups are radically different groups. In particular, 20% of the HE group are simply not comparable with anyone in the no-qualifications group and are dropped from the matching analysis. But, even once we perform matching restricted to the common support, the remaining HE-treated individuals are still so different from the no-qualifications group that the relevant observables characterizing them cannot be adequately controlled for. A considerable degree of imbalance in the X s remains even in the ‘best’ matching specification (in particular, joint balancing of the control variables is rejected at any significance level—see Table 4), revealing how the data simply do not contain enough information for nonparametric identification. Interestingly, when the no-qualifications group is viewed as the treated group to be matched to the larger pool of potential HE comparisons, we obtain a better (though still not acceptable) balancing.

In general, we have found that, the larger the educational gap between the two groups being compared, the more difficult it becomes to balance their characteristics X adequately, this difficulty being further worsened when the potential comparison group is smaller than the treated group. Whereas an OLS specification would have hidden the fundamental non-comparability of these groups, a carefully performed matching estimation could once again highlight the issue of their true comparability and hence the reliability of the results concerning them.

5. Summarizing remarks and conclusions

The aim of this paper has been to review alternative methods and models for the estimation of the effect of education on earnings, and to apply these to a high quality common data source. We have highlighted the importance of the model specification—in particular, the distinction between single-treatment and multiple-treatment models—as well as the importance of allowing for heterogeneous returns—i.e. returns for the same educational qualification that vary across individuals. We have considered four main estimation methods which rely on different identifying assumptions—least squares, IV methods, control function methods and propensity score matching methods. The properties of the estimators were analysed, distinguishing between a single-treatment model and a model where there is a sequence of possible treatments. We argued that the sequential multiple-treatment model is well suited to the education returns formulation, since educational qualification levels in formal schooling tend to be cumulative.

With heterogeneous returns, defining the ‘parameter of interest’ is central. We distinguished four of them: the effect of treatment on the treated individuals, the average treatment effect, the effect of treatment on the non-treated individuals and the LATE. In the homogeneous effects model, these would all be equal, but in the heterogeneous effects model they can differ substantially. Which one is of interest depends on the policy question.

Our application aimed to estimate the wage returns to different educational investments by using the NCDS 1958 birth cohort study for Britain. We argue that this data set is ideally suited for evaluating the effect of education on earnings. There are extensive and commonly administered ability tests at early ages, as well as accurately measured family background and school

type variables, which are all ideal for methods relying on the assumption of selection on observables, notably least squares and matching.

This application has highlighted the following key points.

- (a) Correcting for detailed background variables and differences in ability is important and reduces the return to education at all levels; the basic pre-education information that is available in common data sets would not have been enough to identify gains in an unbiased way.
- (b) The overall returns to educational qualifications at each stage of the educational process remain sizable and significant, even after allowing for heterogeneity in the education response parameters. In particular, we estimate an average return of about 27% for those completing some form of HE compared with anything less. Compared with leaving school at 16 years of age without qualifications, we find that in the population the average return to O-levels is around 18%, to A-levels 24% and to HE 48%.
- (c) We find evidence of heterogeneity in the returns to HE in terms of observables. Furthermore, when we do not allow for such observable heterogeneity in returns, the control function specification points to significant selection on unobserved returns. When we do allow for these interactions, from the control function specification there no longer appears to be any remaining selection on unobserved returns.
- (d) Given the above finding that interactions do matter, an IV approach that is aimed at recovering the average return for the treated individuals calls for a fully interacted IV model, which cannot be estimated precisely with our data. Instead, we recover instrument-related LATEs. We exploit three instruments to glean some information about the extent of variability in returns in the population.
- (e) Overall, matching on detailed early test scores and family background variables appears to perform well for the average return for the treated group in our application to the NCDS data.

Acknowledgements

We thank Bob Butcher, Paul Johnson, Hessel Oosterbeek, Jeff Smith and participants at the Royal Statistical Society Social Statistics Section meeting, the Centre for the Economics of Education 'Returns to education' conference and a seminar at the Policy Studies Institute for comments on earlier drafts and the Joint Editor and the two referees for this journal for their constructive and detailed comments. We also thank Mario Fiorini for research assistance. This research is jointly funded by the Department for Education and Skills Centre for the Economics of Education and the Economic and Social Research Council Centre for the Microeconomic Analysis of Public Policy at the Institute for Fiscal Studies. The usual disclaimer applies.

Appendix A: Classification of educational qualifications

A.1. *The British educational system*

Progression at school beyond the minimum leaving age of 16 years is based on a series of nationally assessed examinations. The wide range of academic and vocational qualifications has been classified into equivalent National Vocational Qualification levels, ranging from level 1 to level 5.

Until 1986, students at 16 years of age had to decide whether to apply for the lower level Certificates of Secondary Education (CSE) option or for the more academically demanding Ordinary level (O-level) route (the top grade (grade 1) achieved on a CSE was considered equivalent to O-level grade C). Whereas most CSE students tended to leave school at the minimum age, students who took O-levels were much more likely to stay on in school. (In 1986 CSEs and O-levels were replaced by General Certificates of

Secondary Education). Those staying on in school can then take Advanced levels (A-levels) at the end of secondary school (age 18 years). A-levels are still the primary route into higher education.

A.2. No qualifications

The no-qualifications group also includes very low level qualifications at National Vocational Qualification level 1 or less, i.e. CSE grade 2–5 qualifications, other business qualifications, other qualifications not specified and Royal Society of Arts level 1 qualifications.

A.3. O-levels or equivalent

The O-levels or equivalent group includes O-levels or CSE grade 1, but also a range of vocational equivalents to these academic school-based qualifications: Royal Society of Arts level 2 and 3; City and Guild operative, craft, intermediate, ordinary or part 1; Joint Industry Board, National Joint Council or other craft or technician certificate.

A.4. A-levels or equivalent

The A-levels or equivalent group includes at least one A-level, but also a range of vocationally equivalent qualifications: City and Guild advanced, final or part 2, or the 3-year or full technological certificate; the insignia award in technology; the Ordinary National Certificate or Ordinary National Diploma, the Scottish National Certificate or Scottish National Diploma, Technician Education Council (TEC) or Business and Technician Education Council (BEC) or the Scottish equivalent SCOTEC and SCOTBEC certificate or diploma.

A.5. Higher education

The HE group includes the Higher National Certificate or Diploma, the Scottish Higher National Certificate or Scottish Higher National Diploma, TEC or BEC, or SCOTEC or SCOTBEC Higher or Higher National Certificate or Diploma, professional qualifications, nursing qualifications including National Nursery Examining Board, polytechnic qualifications, university certificates or diplomas, first degrees, post-graduate diplomas and higher degrees.

A.6. Adjustments to guarantee the sequential nature of the educational variable

Our multiple-treatment estimation method requires sequential educational outcomes; it is thus essential that those who have an A-level or equivalent qualification or HE qualification also have the preceding lower qualifications. This is almost universally true of people who have undertaken an academic route and we impose this in our model. It is, however, not necessarily true for individuals who have undertaken vocational routes; if this is so, we downgrade their qualification by one level to maintain our sequential structure. Specifically: if someone has a first degree or a post-graduate qualification, we assume that they have all the lower qualifications; if someone has one of the other (i.e. vocational) HE qualifications but not an A-level or equivalent qualification, we downgrade their qualification to A-level or equivalent and assign them all the lower qualifications; if someone has an A-level qualification but no O-level qualification, we assign them an O-level qualification; if they have any other A-level equivalent but no O-level equivalent, we downgrade them by one.

References

- Abadie, A. and Imbens, G. (2002) Simple and bias-corrected matching estimators for average treatment effects. *Mimeo*. University of California at Berkeley, Berkeley.
- Angrist, J. (1998) Estimating the labour market impact of voluntary military service using social security data on military applicants. *Econometrica*, **66**, 249–288.
- Angrist, J. and Han, J. (2004) When to control for covariates?: panel asymptotics for estimates of treatment effects. *Rev. Econ. Statist.*, **86**, 58–72.
- Angrist, J. and Imbens, G. (1995) Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *J. Am. Statist. Ass.*, **90**, 431–442.
- Angrist, J., Imbens, G. and Rubin, D. B. (1996) Identification of causal effects using instrumental variables. *J. Am. Statist. Ass.*, **91**, 444–472.
- Angrist, J. and Krueger, A. B. (1991) Does compulsory schooling attendance affect schooling decisions. *Q. J. Econ.*, **106**, 970–1014.

- Angrist, J. and Krueger, A. B. (1992) Estimating the payoff to schooling using the Vietnam-era draft lottery. *Working Paper 4067*. National Bureau of Economic Research, Cambridge.
- Black, D. and Smith, J. (2004) How robust is the evidence on the effects of college quality?: evidence from matching. *J. Econometr.*, **121**, 99–124.
- Blundell, R., Dearden, L., Goodman, A. and Reed, H. (2000) The returns to higher education in Britain: evidence from a British cohort. *Econ. J.*, **110**, F82–F99.
- Blundell, R. and Powell, J. (2003) Endogeneity in nonparametric and semiparametric regression models. In *Advances in Economics and Econometrics* (eds M. Dewatripont, L. Hansen and S. J. Turnovsky). Cambridge: Cambridge University Press.
- Bound, J., Jaeger, D. and Baker, R. (1995) Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *J. Am. Statist. Ass.*, **90**, 443–450.
- Butcher, K. and Case, A. (1994) The effect of sibling sex composition on women's education and earnings. *Q. J. Econ.*, **109**, 531–563.
- Card, D. (1999) The causal effect of education on earnings. In *Handbook of Labor Economics*, vol. 3 (eds O. Ashenfelter and D. Card). Amsterdam: Elsevier.
- Card, D. (2001) Estimating the returns to schooling: progress on some persistent econometric problems. *Econometrica*, **69**, 1127–1160.
- Chevalier, A., Harmon, C. and Walker, I. (2002) Does education raise productivity or just reflect it? *Mimeo*. University of Warwick, Coventry.
- Cochran, W. and Rubin, D. B. (1973) Controlling bias in observational studies. *Sankhya*, **35**, 417–446.
- Dearden, L. (1999a) The effects of families and ability on men's education and earnings in Britain. *Lab. Econ.*, **6**, 551–567.
- Dearden, L. (1999b) Qualifications and earnings in Britain: how reliable are conventional OLS estimates of the returns to education? *Working Paper 99/7*. Institute for Fiscal Studies, London.
- Dearden, L., Ferri, J. and Meghir, C. (2002) The effect of school quality on educational attainment and wages. *Rev. Econ. Statist.*, **84**, 1–20.
- Dehejia, R. H. and Wahba, S. (1999) Causal effects in non-experimental studies: re-evaluating the evaluation of training programmes. *J. Am. Statist. Ass.*, **94**, 1053–1062.
- Fogelman, K. (ed.) (1983) *Growing up in Great Britain: Collected Papers from the National Child Development Study*. London: Macmillan.
- Frölich, M. (2004) Finite sample properties of propensity-score matching and weighting estimators. *Rev. Econ. Statist.*, **86**, 77–90.
- Garen, J. (1984) The returns to schooling: a selectivity bias approach with a continuous choice variable. *Econometrica*, **52**, 1199–1218.
- Gosling, A., Machin, S. and Meghir, C. (2000) The changing distribution of male wages, 1966–93. *Rev. Econ. Stud.*, **67**, 635–666.
- Griliches, Z. (1977) Estimating the returns to schooling: some econometric problems. *Econometrica*, **45**, 1–22.
- Harmon, C. and Walker, I. (1995) Estimates of the economic return to schooling for the UK. *Am. Econ. Rev.*, **85**, 1278–1286.
- Heckman, J. J. (1979) Sample selection bias as a specification error. *Econometrica*, **47**, 153–161.
- Heckman, J. J., Ichimura, H. and Todd, P. (1997) Matching as an econometric evaluation estimator: evidence from evaluating a job training programme. *Rev. Econ. Stud.*, **64**, 605–654.
- Heckman, J. J., Ichimura, H. and Todd, P. (1998) Matching as an econometric evaluation estimator. *Rev. Econ. Stud.*, **65**, 261–294.
- Heckman, J. J., Ichimura, H., Smith, J. and Todd, P. (1998) Characterizing selection bias using experimental data. *Econometrica*, **66**, 1017–1098.
- Heckman, J. J., LaLonde, R. and Smith, J. (1999) The economics and econometrics of active labor market programs. In *Handbook of Labor Economics*, vol. 3 (eds O. Ashenfelter and D. Card). Amsterdam: Elsevier.
- Heckman, J. J. and Navarro-Lozano, L. (2004) Using matching, instrumental variables, and control functions to estimate economic choice models. *Rev. Econ. Statist.*, **86**, 30–57.
- Heckman, J. J. and Robb, R. (1985) Alternative methods for evaluating the impact of interventions. In *Longitudinal Analysis of Labour Market Data* (eds J. J. Heckman and B. Singer). Cambridge: Cambridge University Press.
- Heckman, J. J., Smith, J. and Clements, N. (1997) Making the most out of program evaluations and social experiments: accounting for heterogeneity in program impacts. *Rev. Econ. Stud.*, **64**, 487–536.
- Heckman, J. and Vytlačil, E. (2000) Identifying the role of cognitive ability in explaining the level of and change in the return to schooling. *Working Paper W7820*. National Bureau of Economic Research, Cambridge.
- Holland, P. W. (1986) Rejoinder. *J. Am. Statist. Ass.*, **81**, 968–970.
- Ichino, A. and Winter-Ebmer, R. (1999) Lower and upper bounds of returns to schooling: an exercise in IV estimation with different instruments. *Eur. Econ. Rev.*, **43**, 889–901.
- Imbens, G. (2000) The role of propensity score in estimating dose-response functions. *Biometrika*, **87**, 706–710.
- Imbens, G. (2004) Semiparametric estimation of average treatment effects under exogeneity: a review. *Rev. Econ. Statist.*, **86**, 4–29.

- Imbens, G. and Angrist, J. (1994) Identification and estimation of local average treatment effects. *Econometrica*, **62**, 467–476.
- Kane, T., Rouse, C. E. and Staiger, D. (1999) Estimating the returns to schooling when schooling is misreported. *Working Paper 7235*. National Bureau of Economic Research, Cambridge.
- Lechner, M. (2001a) Identification and estimation of causal effects of multiple treatments under the conditional independence assumption. In *Econometric Evaluation of Labour Market Policies* (eds M. Lechner and F. Pfeiffer). Heidelberg: Physica.
- Lechner, M. (2001b) A note on the common support problem in applied evaluation studies. *Discussion Paper 2001-01*. Department of Economics, University of St Gallen, St Gallen.
- Leuven, E. and Sianesi, B. (2003) PSMATCH2: Stata module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing. University of Amsterdam and Institute for Fiscal Studies, London. (Available from <http://ideas.repec.org/c/boc/bocode/s432001.html>.)
- McCulloch, A. and Joshi, H. (2002) Child development and family resources: an exploration of evidence from the second generation of the 1958 birth cohort. *J. Popln Econ.*, **15**, 283–304.
- McIntosh, S. (2004) Further analysis of the returns to academic and vocational qualifications. *Discussion Paper 35*. Centre for Economics of Education, London.
- Meghir, C. and Palme, M. (2000) Estimating the effect of schooling on earnings using a social experiment. *Working Paper 99/12*. Institute for Fiscal Studies, London.
- Powell, J. (1994) Estimation of semiparametric models. In *Handbook of Econometrics*, vol. 4 (eds R. F. Engle and D. L. McFadden). Amsterdam: Elsevier.
- Rosenbaum, P. R. and Rubin, D. B. (1983) The central role of the propensity score in observational studies for causal effects. *Biometrika*, **70**, 41–55.
- Rosenbaum, P. R. and Rubin, D. B. (1985) Constructing a comparison group using multivariate matched sampling methods that incorporate the propensity score. *Am. Statistn*, **39**, 33–38.
- Rubin, D. B. (1980) Discussion of “Randomisation analysis of experimental data in the Fisher randomisation test” by Basu. *J. Am. Statist. Ass.*, **75**, 591–593.
- Rubin, D. B. (1986) Discussion of “Statistics and causal inference” by Holland. *J. Am. Statist. Ass.*, **81**, 961–962.
- Schmitt, J. (1995) The changing structure of male earnings in Britain, 1974–88. In *Changes and Differences in Wage Structures* (eds R. Freeman and L. Katz). Chicago: University of Chicago Press.
- Shea, J. (1997) Instrument relevance in multivariate linear models: a simple measure. *Rev. Econ. Statist.*, **79**, 348–352.
- Sianesi, B. (2002) Essays on the evaluation of social programmes and educational qualifications. *PhD Thesis*. University College London, London.
- Sianesi, B. (2004) An evaluation of the Swedish system of active labour market programmes in the 1990s. *Rev. Econ. Statist.*, **86**, 133–155.
- Smith, J. and Todd, P. (2004) Does matching overcome LaLonde’s critique of nonexperimental estimators? *J. Econometr.*, **125**, 305–353.
- Staiger, D. and Stock, J. H. (1997) Instrumental variables regressions with weak instruments. *Econometrica*, **65**, 557–586.
- Wooldridge, J. (1997) On two stage least squares estimation of the average treatment effect in a random coefficient model. *Econ. Lett.*, **56**, 129–133.
- Zhao, Z. (2004) Using matching to estimate treatment effects: data requirements, matching metrics, and Monte Carlo evidence. *Rev. Econ. Statist.*, **86**, 91–107.