Estimating the Returns to Education: Models, Methods and Results

Richard Blundell

Lorraine Dearden

Barbara Sianesi

October 2001

Published by
Centre for the Economics of Education
London School of Economics
Houghton Street
London WC2A 2AE

October 2001
Individual copy price: £5

The Centre for the Economics of Education is an independent research centre funded by the Department for Education and Skills. The views expressed in this work are those of the author and do not reflect the views of the DfES. All errors and omissions remain the authors.

Estimating the Returns to Education: Models, Methods and Results

Richard Blundell

Lorraine Dearden

Barbara Sianesi

Acknowledgments

Richard Blundell is a Professor of Economics at the Department of Economics, University College London, Research Director at the Institute for Fiscal Studies and a Research Associate at the Centre for the Economics of Education. Lorraine Dearden is Director of the Education, Employment and Evaluation Research sector, Institute of Fiscal Studies. Barbara Sianesi is a Research Economist at the Institute for Fiscal Studies.

The authors would like to thank Bob Butcher, Paul Johnson and participants at the RSS meeting for comments on earlier drafts. Thanks also to thank Mario Fiorini for research assistance.

Estimating the Returns to Education: Models, Methods and Results*

Richard Blundell, Lorraine Dearden and Barbara Sianesi

University College London and Institute for Fiscal Studies

October 2001

Abstract

This paper reviews appropriate non-experimental methods and microeconometric models for recovering the returns to education using individual data. Three estimators are considered: matching methods, instrumental variable methods and control function methods. The properties of these methods are investigated for models with multiple treatments and heterogeneous returns. Data from the British 1958 NCDS birth cohort is used to estimate returns to schooling and to illustrate the sensitivity of different estimators to model specification and data availability.

1 Introduction

There are at least three distinct ways of defining the 'returns to education': (a) the private return, (b) the social return and (c) the labour productivity return. The first of these is made up of the costs and benefits to the individual and is clearly net of any transfers from the state and any taxes paid. The second definition highlights any externalities or spill-over effects and includes transfers and taxes. The final definition simply relates to the gross increase in labour productivity (or growth). A key component of each of these measures is the impact of education on earnings. This is perhaps the aspect of returns to education measurement where statistical methods have been most developed and most fruitfully deployed and is the central focus of this paper.

With extensive data available over time and individuals on schooling and on earnings, the measurement of the education effect on earnings is one area where we might expect agreement. However, a casual look through the literature on the impact of education on earnings reveals a wide range of estimates and an equally wide range of empirical approaches that have been adopted to estimate the return. So why do the estimates vary so widely and what is the most appropriate empirical method to adopt? The answer to these two questions provides the central motivation for this paper. It is illustrated using the sample of men from the NCDS Birth Cohort data for the UK. This data source provides a uniquely rich source of non-experimental data on family background, educational attainment and earnings. We argue that it is ideally suited to analyse statistical methods for the measurement of the effect of education on earnings.

The appropriate statistical method to adopt will depend, in a rather obvious way, on the chosen model for the relationship between education and earnings. We distinguish two broad characterisations of this specification. The first relates to the measurement of education. In particular whether we can summarize education, or human capital more generally, in a single measure - years of schooling, for example. This is commonly referred to as a one factor model. It is a restrictive framework since it assumes that, as returns to education change over time, it is only the single aggregate that matters and there are no differential trends in returns for different education levels. It is convenient though since we can simply include a single education measure in an earnings equation. The alternative to this is the multiple factor model where different educational levels have separate effects on earnings.

We will refer to different education levels as different treatments borrowing a common notation from the evaluation literature. A single treatment specification refers to the impact of a specific educational level - such as gaining a qualification. A multiple treatment effect model will distinguish the impact of many different education levels. In general the multiple treatment - multiple factor model would seem a more attractive framework since we will typically be interested in a wide range of education levels with very different returns. However, we will also consider models with a single discrete treatment such as the impact of a specific qualification and models with a single overall education level such as years of schooling.

The second characterisation relates to the distinction between heterogeneous and homogeneous returns. In simple terms - whether the response coefficient on the education variable(s) in the earnings equation is allowed to differ across individuals. To allow this to happen according to observables is a relatively straightforward extension of the homogeneous model, but to allow the heterogeneity to be unobservable completely changes the interpretation and the properties of many common estimation approaches. We begin in Section 2 with this distinction between model specifications and use it to define parameters of interest in the earnings education relationship.

Even where there is agreement on the model specification there are alternative statistical methods which can be adopted. With experimental data the standard comparison of control and treatment group recovers an estimate of the average effect in the population under the assumption that the controls are unaffected by the treatment. Although in some studies of training, experimental design is possible and growing in popularity, for large reforms to schooling, and for measuring the impact of existing educational systems, nonexperimental methods are essential. There are broadly two nonexperimental methods: those that attempt to control for 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 match on these observed variables. The implementation and the properties of these alternative methods differ according to whether the model is one of heterogeneous response and whether schooling is represented through a single or multiple measure. The different properties of these estimators and the drawbacks to each method are discussed in Section 3.

The various models and non-experimental estimators are then compared in Section 4 and 5 using the British NCDS data. In particular, we consider the returns to education for three levels: 1) leaving after completing O levels or its vocational equivalent 2) leaving after completing A levels or its vocational equivalent 3) undertake some form of higher education (including sub-degree level HE).

We only present results for men so as to conserve space and to focus on the earnings effect versus the employment effect. Our results show significant returns to basic and higher qualifications in education, even after controlling for detailed family background and ability test measures. They also highlight two important features: (a) the returns to education on earnings are sensitive to the inclusion of controls. In particular we find that controlling for ability test scores at an early age significantly reduces the returns by a factor of one third. (b) there is strong evidence of heterogeneity of returns with quite high returns found among subsets of those who do not stay on for higher qualifications at school. Although on average the returns among those that stay on for higher qualifications are higher than for those who do not.

2 The Earnings-Education Relationship: Alternative Models

This section highlights two central aspects in the empirical investigation of the earnings return to human capital investments. First among these is the distinction between the *homogeneous returns* and *heterogeneous returns* model. In the homogeneous returns model the rate of return to gross earnings of a particular human investment is the same for all individuals. Growing statistical evidence and causal empiricism suggests that the homogeneous returns restriction is unwarranted.

The second aspect is to distinguish between the *one factor* and *multiple factor* models of human capital. In the one factor model all schooling can be thought of as an investment in a single homogeneous construct called human capital. Each additional unit has the same return. An example of an empirical model that is both one factor and homogeneous returns is the popular linear regression equation - log earnings regressed on years of schooling. The constant parameter on the schooling variable is equivalent to homogeneous returns and the use of years of schooling as a single measure of schooling is equivalent to a single measure of human capital.

2.1 Earnings and Education in a Homogenous Returns Framework

For each individual $i = 1, ..., n$, we let y_i represents their earnings or hourly wage opportunities in work. To begin with we will assume that we are measuring earnings at one point in time for a sample of individuals who have completed formal schooling. A good illustration to keep in mind is from the British cohort studies where a single cohort is followed through education and employment and sampled at specific intervals usually several years apart. We measure their earnings when they are 33 years of age and ask: what is the impact on earnings at age 33 of different schooling outcomes?

In the "one factor" human capital model it will always be possible to aggregate schooling into a single measure S_i - say, years of schooling. For example, we may write

$$\ln y_i = \alpha_i + \beta S_i + \varepsilon_i \tag{1}$$

where α_i represents differing relative levels of earnings across individuals for any given level of schooling and β measures the marginal return to schooling level S_i in terms of the particular definition of earnings y_i. The "error" term ε_i is added to capture measurement error in earnings.[1]

Since educational choices and educational levels are likely to differ according to productivity (or expected earnings levels more generally), S_i is very likely to be positively correlated with α_i and this in turn will induce an upward bias in the simple least squares estimation of β. However, if S_i is measured with error there will be some off-setting attenuation bias and this trade-off was at the heart of the early studies on measuring gross private returns (see Griliches (1977) and Card (1999), for example). We will return to these estimation issues in more detail below.

As mentioned above (1) assumes both homogeneous returns (the common β across all i) and one factor human capital (the single measure of schooling S_i). Although popular this seems unduly restrictive. In order to cover a fairly flexible representation of schooling we will consider the case of a finite set of schooling levels. For example, in the application to the NCDS we refer to a number of specific discrete educational levels - as obtaining a qualification, obtaining an A level, or undertaking higher education. Using notation borrowed from the experimental literature we will refer to these as *multiple treatments.* These will typically be defined in some natural sequence of binary indicator variables:

[1] Measurement error in the schooling or eduction variable S_i is also likely to be important and will be discussed in terms of the alternative approaches to estimation.

$S_{1i} = 1$ if the individual completed the first stage of schooling, $S_{2i} = 1$ if the next stage in the sequence is completed etc. For completeness $S_{0i} = 1$ would represent the base educational level. For example in the UK context $S_{1i} = 1$ may refer to staying on after the minimum school leaving age, $S_{2i} = 1$ might represent achieving at least one A-level, $S_{2i} = 1$ to achieving a first degree etc. Write the exhaustive set of J treatments (schooling levels) under examination as S_{1i}, S_{2i}, ..,S_{Ji}. In this case (1) might be adapted to become

$$\ln y_i = \alpha_i + \beta_1 S_{1i} + \beta_2 S_{2i} + .. + \beta_J S_{Ji} + \varepsilon_i \tag{2}$$

where the β_1, $\beta_2 ... \beta_J$. now measure the marginal impact of a higher level of schooling for some J distinct levels. Still the returns are homogeneous across individuals but (2) can be seen to relax the "one factor" assumption and allow different schooling levels to have quite different impact on earnings.

Of course, one can imagine a finer sequence and also possible set of non-sequential outcomes. All the methods discussed below are easily extended to more complicated situations but will typically require more demanding data requirements and modelling assumptions to estimate the "causal" impact on earnings. Indeed, to begin the discussion of the heterogeneous model we will begin with a single treatment model and study the "causal" impact of a single type of schooling level.

Before moving on to the more general models and the alternative methods of estimation, it is worth pointing out that each of these equations will typically be specific to a particular time period and location. For example, if (2) refers to the impact of education levels on the earnings of British men aged 33 in 1991, it will be unlikely to be stable across time periods and countries. The returns will depend on the earnings set in the labour market and will in turn depend on the demand and supply of individuals with these differing human capital attributes. This point, although quite obvious, is often misunderstood in the context of predicting returns to education.

2.2 The Heterogeneous Returns Model

The heterogeneity we are focussing on here is unobserved heterogeneity across individuals in the response parameter β. Consider the single treatment model where we let schooling S_{1i} for individual i be defined as a binary indicator variable representing the successful achievement of a particular education level - such as obtaining a qualification, obtaining an A level, or undertaking higher education, for example. A completely general relationship between this level of education and earnings in this *single discrete treatment heterogeneous returns model* is then written

$$\ln y_i = \alpha_i + \beta_i S_{1i} + \varepsilon_i \tag{3}$$

where α_i and β_i can be thought of as random coefficients representing the heterogeneous relationship between educational qualification S_{1i} and earnings. Typically we would assume the α_i and β_i have a finite population mean and variance. Below the population means are labelled α_0 and β_0 respectively.

Despite the preponderance of the homogeneous returns model in the early literature, the recent focus has been on the heterogeneous returns model (3)[2]. This raises the immediate question in this model: what is the parameter of interest. Is it the average of the $\beta_i's$? If so what average? Is it the average in the population whether or not level S_{1i} is achieved, β_0 - the *average treatment effect*, or the average among those who individuals actually observed with $S_{1i} = 1$, β_T - the *average treatment on the treated*? In some cases a particular estimation method will recover a *local average treatment effect,* measuring the impact of $S_{1i} = 1$ on an even smaller subgroup of individuals. We discuss all these in greater detail in the next section.

[2] See for example the papers by Heckman, J., Smith, J. And N. Clements, (1997), Dearden (1999a) and (1999b) and Blundell, Dearden, Goodman and Reed (2000).

One interpretation of β_i is as a heterogeneous "return" to schooling level S_{1i} for individual i since it measures the marginal proportional impact of this level of education on earnings for individual i. Again this measure of returns is a gross private measure since it ignores all costs of education and also taxes paid on gross earnings y_i. Note that in the "homogeneous" returns model β_i is constant across all individuals. Even so, in this homogeneous returns model α_i is allowed to vary across i to capture the differing productivities (or abilities), and differing general levels of earnings, across individuals with the same education levels.

It should be pointed out that this model, and the others to be discussed below, can be readily generalised to allow for *observable heterogeneity* in both α_i and β_i. For example, suppose there are a set of observed covariates X_i (e.g. early test scores, demographic variables, aspects of the local labor market). The α and β parameters can be made to depend on these in a quite arbitrary way. If they are assumed to depend on X_i in a linear fashion then the levels of X_i and the interactions of X_i with the education variable S_{1i} will enter the regression specification. The precise form chosen will depend on the richness of the data set and the particular problem at hand[3]. But in what follows we shall always assume that such levels and interactions are included in the specification. Indeed, in the general nonparametric matching method described below a quite general form of interaction is allowed. For the most part we will assume that such observed heterogeneity terms are included, even if this is not explicitly stated in the discussion of the properties of the various alternative estimators described below.

As we saw in the homogeneous model (1), the dependence of the schooling level S_i on the unobserved "ability" component α_i is critical in understanding the bias from standard least squares estimation. An additional central issue in determining the properties of standard econometric estimators in the hetero-

[3] This was done in Dearden (1999a) using the same NCDS data that is used in this paper.

geneous effects model is whether or not schooling choice S_{1i} depends on the unobservable determinants of the individuals's marginal return from schooling β_i. If β_i were known when the individual makes his or her educational choices then it would seem sensible to assume that choices will - in part, at least, reflect the return to earnings of that choice. But as mentioned before β_i is likely to vary over time and will depend on the relative levels of demand and supply, so the dependence of schooling choices on marginal returns is not clear-cut. Some persistence in returns is however likely and so some correlation would seem more likely than not.

The discussion of heterogeneous returns will extend easily to the *multiple treatment model* (2). Writing the exhaustive set of J treatments (schooling levels) under examination as S_{1i}, S_{2i}, ..,S_{Ji}. The heterogeneous returns model is then

$$\ln y_i = \alpha_i + \beta_{1i}S_{1i} + \beta_{2i}S_{2i} + .. + \beta_{Ji}S_{Ji} + \varepsilon_i. \qquad (4)$$

We will also want to discuss the *one factor model* in which S_i enters as a single continuous variable

$$\ln y_i = \alpha_i + \beta_i S_i + \varepsilon_i. \qquad (5)$$

In fact, the three basic specifications (3), (2) and (5) will form the main alternatives considered here. The single discrete treatment case (3) being the baseline specification.

3 The Earnings-Education Relationship: Alternative Methods

The aim here is to investigate the properties of alternative estimation methods for each of the model specifications considered above. The alternative methods we consider fall into three broad classes: the instrumental variable method, the control function method and the method of matching. The first two require some excluded instrument which determines education choices but not earnings while the matching method requires an extensive set of observable characteristics on which to match. All place strong demands on data.

The initial setting for this discussion will be based on the biases that occur from the simple application of ordinary least squares to the estimation of each of the model specifications described in the previous section. As mentioned the primary model specification will be the single discrete treatment heterogeneous returns model (3) but the extension to the multiple treatment model (4) will also be considered and so will the specific issues that occur in the one factor "years of schooling" specification (5). In each of these the complications that are engendered by allowing the return parameter β to be heterogeneous will be central to the discussion.

It will also be useful to relate this choice of alternative methods to the evaluation and the selection literature (see in particular Heckman and Robb (1985) and Heckman, LaLonde and Smith (1999)). The treatment effect parameters referred to above already borrow heavily from this literature. It is easy to see that even the estimation problem in the education returns framework is synonymous with the construction of a counter-factual in the evaluation literature. Indeed, some of the more recent developments in the in the returns literature, for example those which use matching estimators or social experiments, relate explicitly to similar approaches in evaluation.

It is also worth pointing out that we will not be looking at general spillover effects or general equilibrium effects of education. In the statistical evaluation literature this relates to the stable unit-treatment value assumption (SUTVA). This assumption requires that an individual's potential outcomes depend only on his own schooling participation, not on the schooling choice of other individuals in the population (thus ruling out cross-effects or general equilibrium effects) and that the education level chosen by an individual does not depend on the schooling decisions of others (e.g. thus excluding peer effects in educational choices).

To illustrate the importance of constructing the counterfactual, consider the single discrete treatment model (3). For any individual i in the set of individuals that receive $S_{1i} = 1$ ($i \in \{S_{1i} = 1\}$) the earnings outcome is

$$\ln y_i^1 = \alpha_i + \beta_i + \varepsilon_i \text{ for } i \in \{S_{1i} = 1\} \tag{6}$$

where the superscript 1 refers to the case where individual i receives treatment $S_{1i} = 1$. Whereas if the same individual were not to receive this education level their earnings outcome would be

$$\ln y_i^0 = \alpha_i + \varepsilon_i \text{ for } i \in \{S_{1i} = 1\}. \tag{7}$$

now the superscript 0 refers to the counterfactual earnings of an individual i for whom $S_{1i} = 1$ in the observed data. Suppose that in a random sample of size n there are n_1 individuals for whom $S_{1i} = 1$. If we could observe both outcomes for all individuals for whom $S_{1i} = 1$ then the average

$$\sum_{i \in S_{1i}=1} \frac{\ln y_i^1 - \ln y_i^0}{n_1} \tag{8}$$

would be a consistent estimate of the *average treatment on the treated* effect β_T. To recover the *average treatment effect* β_0 across the whole population under consideration further requires the counterfactual for the group for which $S_{1i} = 0$.

The focus here will be on non-experimental approaches. Of course, in a randomized experiment the control group is chosen independently of the α_i, β_i and ε_i by design. Consequently the average treatment effect can be measured directly from a comparison of the control group and the treatment group. This is the case even in the most general heterogeneous returns model described above and whatever the underlying relationship between the education level variables S_{ji} etc., and the heterogeneous ability and returns parameters α_i and β_i parameters respectively. A truly randomised education experiment would induce full independence between these individual heterogeneity parameters and the education outcomes. However, pure education or schooling experiments are very rare. It is difficult to persuade parents or the students themselves of the virtues of being randomised out of an education programme - except for rather minor programmes. Our application will be to the main stages of educational level in the UK and randomised assignment is unavailable. Even if it were, individuals enrolled may drop out - and systematically too.

However, the pure randomised experiment is useful as a basis for comparison. "Natural" social experiments are more common. This is the case where some educational rule or qualification level (say minimum schooling leaving age) is exogenously changed for one group but not another. Provided the groups are representative samples from the population then this simple comparison can recover a parameter of interest like the average treatment effect. Where the samples differ in their ability levels or other characteristics it may still be possible to recover an average effect for those who experience the change in rules.

Many alternative methods are available and all have been widely used. Twins and sibling data can help, an exogenous change in S_i that only effects a subsample of the target group is also useful. The latter relates to instrumental variable methods more generally, where some variable (or transformation of the data) has to be found that can vary S_i independently of the heterogeneity terms. An alternative to using instruments is the matching method. This method seeks to purge the relationship between schooling and earnings of any important observed heterogeneity that would lead to bias, by matching individuals with and without the schooling level according to some observed characteristic.

3.1 Least Squares

Consider rewriting the heterogeneous single treatment model (3) as

$$\ln y_i = \alpha_0 + \beta_0 S_{1i} + (\alpha_i - \alpha_0) + (\beta_i - \beta_0) S_{1i} + \varepsilon_i \tag{9}$$

where α_0 and β_0 are the population means of α_i and β_i. This population may be defined as all those individuals entering schooling at a particular date. In this case a parameter of interest will be β_0 itself which measures the average returns to achieving education level S_{1i} in this population. For example, in the British context $S_{1i} = 1$ may refer to those staying on after the minimum school leaving age, or it may refer to those going on to a higher degree.

Gathering the unobservables together we have

$$\ln y_i = \alpha_0 + \beta_0 S_{1i} + u_i \tag{10}$$

with

$$u_i \equiv (\alpha_i - \alpha_0) + (\beta_i - \beta_0) S_{1i} + \varepsilon_i$$

The least squares regression of log earnings on schooling produces a biased estimator of β_0 due to correlation between S_{1i} and u_i. There are three sources of such bias:

(i) Ability bias: this occurs due to the likely correlation between S_{1i} and the $(\alpha_i - \alpha_0)$ term. A positive correlation inducing an upward bias in the estimated return.

(ii) Returns Bias: this occurs when the marginal returns $(\beta_i - \beta_0)$ themselves are correlated with the schooling choice S_{1i}. The direction of this bias is less clear and will depend on the average returns among the sub-population of those with S_{1i}. Indeed, if ability bias is negligible (and the remaining ability heterogeneity is unrelated to the unobserved return) and returns bias is the only remaining bias present, then the least squares coefficient estimate will recover the average returns in the sample of those with $S_{1i} = 1$, that is the average treatment on the treated.

(iii) Measurement Error Bias: this refers to measurement error in the schooling variable S_{1i}. This may be due to random misclassification error. As usual, measurement error of this kind will induce attenuation bias in the regression coefficient and an under-estimate of the returns parameter. For the purposes of much of the discussion we can redefine ε_i to include measurement error in the schooling variables(s).

In the homogeneous returns model the second 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 (1). Indeed, there is some evidence of a balancing of biases (see Card (1999), for example), in which case OLS fortuitously consistently estimates the return coefficient β.

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 omitted factors. Indeed, the method of matching, described below, takes this one step further by trying to eliminate the imbalance of observables by matching observations with similar covariates. One simple recommendation in the use of least squares seems worth following - add in a rich set of controls and try to find separate measures of the education variable that may not suffer from the same measurement error. The rich set of controls may help reduce the ability bias and the second measure of schooling may be used to purge the measurement error. But, in the end, a comparison with the alternative methods: instrumental variables, control functions and matching is always helpful in assessing how to interpret a least squares estimate of education returns.

3.2 Instrumental Variable Methods

The Instrumental Variable (IV) estimator seems a natural method to turn to in estimating returns - at least in the homogeneous returns model. The biases we have discussed in the case of least squares all stem 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 unobservable ability, heterogeneity and measurement error terms, then surely a consistent estimator of the returns is achievable. This turns out to be true in the homogeneous returns model but not, except for certain special cases described below, for the heterogeneous returns model.

Even in the homogeneous returns model, finding a suitable instrument is not easy. However, social and natural experiments can be useful - and many such instruments have been used. Alternatively, parental background variables are often chosen. But to be useful they must satisfy the Instrumental Variable criteria of being correlated with the schooling choice and correctly excluded from the earnings equation.

To more formally investigate the properties of the IV estimator, define an instrumental variable Z_i and assume that it satisfies the orthogonality conditions[4]:

IV: A1: $E[(\alpha_i - \alpha_0)|Z_i] = 0$

IV: A2: $E[(\beta_i - \beta_0)|Z_i] = 0$

IV: A3: $E[\varepsilon_i|Z_i] = 0$

Additionally assume that the instrumental variable is related to S_{1i} through

IV: A4: $E[S_{1i}|Z_i] = Z_i'\pi$

[4] In fact the IV estimator we will discuss will typically only require the weaker covariance restrictions $cov(\alpha_i Z_i) = cov(\beta_i Z_i) = cov(\varepsilon_i Z_i) = 0$ rather than these mean independence assumptions.

where π is a finite vector of unknown reduced form coefficients.

To discuss the general properties of the IV estimator in the heterogeneous returns model consider the conditional expectation of (9) under assumptions **IV: A1 - A4**

$$E[\ln y_i | Z_i] = \alpha_0 + \beta_0 Z_i' \pi + E[(\beta_i - \beta_0) S_{1i} | Z_i]. \tag{11}$$

Note that **IV: A4** implies that π can be estimated consistently from the least squares regression of S_{1i} on Z_i - the first stage, reduced form regression. There is nothing in assumptions **IV: A1 - A4** that makes the final term in (11) disappear. Further assumptions are required. Indeed, even **IV: A4** is controversial since in this specification S_{1i} is discrete and this assumption is better suited to the one factor "years of schooling" model. Indeed, in the homogeneous "one factor" model β_i is constant across i and the latter term is zero by definition. Consequently, instrumental variable estimation can produce a consistent estimator of β in this case.

One way to interpret the IV estimator is as the (weighted) least squares estimator for the transformed regression model

$$Z_i \ln y_i = Z_i \alpha_i + \beta_i Z_i S_{1i} + Z_i \varepsilon_i \tag{12}$$

with the weights depending on the sample covariance matrix of $Z_i \varepsilon_i$. In the case where there is a single instrument the IV estimator reduces to

$$\widehat{\beta}_{IV} = \frac{cov(\ln y_i, Z_i)}{cov(S_i, Z_i)} \tag{13}$$

Consequently, given the IV assumptions IV: A1 - A4, an IV transformation Z_i that eliminates α_i but leaves $\beta_i S_{1i}$ unaffected, will estimate the average β_i among those individuals for whom $S_{1i} = 1$. This is the treatment on the treated parameter $\widehat{\beta}_{IV} = \beta_T$. A prime example of such a transformation is a differencing instrument. There are three such differencing transformations commonly used.

The first is the *difference in differences* estimator. This compares the group of individuals with $S_{1i} = 1$ to the group with $S_{1i} = 0$ before and after the treatment S_{1i} occurs. This transformation sweeps away any common individual component (α_i, for example) and just picks out the average of β_i among those individuals with $S_{1i} = 1$. Precisely the treatment on the treated parameter. The difference in differences estimator is ideally suited for training programmes for which there is an earnings observation before and after training but less applicable in the schooling/formal education evaluation problem where a "before" observation of earnings is very unlikely to be available, since for most individuals formal education is completed before labour market participation.

A second such differencing instrumental variable estimator is to use a *sample of twins*[5]. The contrast in this case is between individuals within pairs of twins. Differencing within twins eliminates common gene determined ability or family effects. These are often what is being represented by α_i. In the homogeneous returns model this perfectly identifies β. However, it only uses information on twins who have different S_{1i} outcomes. This may only be a small subset of twins. Typically twins separated at birth are used and this tends to increase the probability of different values of S_{1i} within any pair. In the heterogeneous returns model the twin difference estimator recovers the average marginal return on those twins with different schooling outcomes. This may be a representative sample of the population but it may not. In general what is estimated strictly reflects the local average of the returns among the sample of twins who experience both outcomes.

Finally, more general groups can be compared. For example we may compare the outcomes among two groups that have a similar distribution of abilities but who, from some exogenous reform, experience different schooling outcomes[6].

[5] See for example Ashenfelter and Krueger (1994), Altonji and Dunn (1996), Ashenfelter and Zimmerman (1993), Behrman et. al. (1996), Bonjour et. al. (2000).

[6] For example see the papers by Angrist and Krueger (1991) and (1992), Butcher and Case

A classic example of this is the comparison of adjacent cohorts one of which experiences a school reform (say a change in the minimum school leaving age) and the other who does not[7]. Again in the homogenous treatment effects model this can be used to estimate β, but in the heterogeneous model it will estimate the average of returns among those induced to take more schooling by the reform. This is the Local Average Treatment Effect and will not in general equal the average returns among those with S_{1i} - the treatment on the treated parameter, or the average in the population - the average treatment effect. The way in which the IV estimator for the heterogeneous returns model is related to a local average of the returns is investigated further below. First we consider some special cases.

Even where the IV estimator does produce a consistent estimate of β there remains the issue of efficiency and of weak instruments. Efficiency concerns the imprecision 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 zero. In this case IV will tend to the biased OLS estimator even in very large samples (see Bound et.al. (1995) and Staiger and Stock (1997)).

3.2.1 IV in the Heterogeneous "One Factor" Model

Consider the one factor or "years of schooling" model (5). This is a case where assumption IV A4 would seem plausible and note that this assumption implies:

$$S_i = Z_i'\pi + v_i \text{ where } E(v_i|Z_i) = 0. \tag{14}$$

and a consistent estimate of π can be obtained from OLS on the reduced form Now assume

(1994), Harmon and Walker (1995), Meghir and Palme (2000).

[7] This was the reform exploited by Harmon and Walker (1995).

IV A5: $v_i = \rho_{v\beta}(\beta_i - \beta_0) + \eta_i$, with $E[(\beta_i - \beta_0)^2] = \sigma^2_\beta$.

This assumption implies

$$E[(\beta_i - \beta_0)S_i|Z_i] = \rho_{v\beta}\sigma^2_\beta$$

so that (11) becomes

$$\begin{aligned} E[\ln y_i|Z_i] &= \alpha_0 + \beta_0 Z_i'\pi + \rho_{v\beta}\sigma^2_\beta \\ &= \widetilde{\alpha}_0 + \beta_0 Z_i'\pi \qquad (15) \end{aligned}$$

so that the IV estimator will, under these assumptions, consistently estimate the average return β_0 but not the intercept.

Note that this is very specific to the continuous schooling measure S_i in the one factor model, since then the additively separable model for (14) is a reasonable specification for the reduced form. Even so the homoscedasticity assumption imbedded in IV A4 is strong. A useful generalisation of this IV estimator for this specification occurs in the Control Variable approach discussed below.

3.2.2 IV in the Homogeneous One Factor Model

By definition this is a specification in which β_i is constant for all i. Consequently $\sigma^2_\beta = 0$ and (11) becomes

$$E[\ln y_i|Z_i] = \alpha_0 + \beta_0 Z_i'\pi \qquad (16)$$

so that IV consistently estimates the α_0 and β_0 parameters.

Note that the exact same IV estimator can be computed from a regression of log earnings on schooling S_i including the reduced form error v_i as an additional regressor

$$\ln y_i = \alpha_0 + \beta_0 S_i + \rho_{uv} v_i + \varsigma_i \text{ where } E[\varsigma_i|S_i, v_i] = 0 \qquad (17)$$

Replacing v_i by

$$\widehat{v}_i = S_i - Z_i'\widehat{\pi}$$

i.e. the residual from the reduced form, preserves the correspondence with IV. This augmented regression framework for IV is popular for testing the exogeneity assumption ($H_0 : \rho_{uv} = 0$) and generalises to binary choice and censored regression settings (see Smith and Blundell (1986)).

The estimation of α and β by the inclusion of v_i in this homogeneous returns specification is also exactly equivalent to the control function approach . As we will show below, this analogy between IV and control function breaks down outside the one factor homogeneous returns model.

3.2.3 IV in the Heterogeneous Single Treatment Model: Estimating the Local Average Treatment Effect

Even in the general heterogeneous returns model with a single treatment (3), it is still possible to provide an interesting interpretation of the IV estimator even if it does not estimate the average treatment on the treated or average treatment parameter. The interpretation of IV in this model specification was precisely the motivation for the Local Average Treatment Effect of Imbens and Angrist (1994).

Suppose there is a single discrete binary instrument $Z_i = \{0, 1\}$. For example, a discrete change in some educational ruling that is correlated with the schooling level S_i in the population. There will be four subgroups of individuals, one of these is of particular interest and is made up of those individuals who are seen with education level $S_{1i} = 1$ after the rule change ($Z_i = 1$) but who would not have had this level of schooling in the absence of the rule change ($Z_i = 0$). This is the group induced to change behavior by the instrument. 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

LATE: A1 For all i ether $[D_{1i} \geq D_{0i}]$ or $[D_{1i} \leq D_{0i}]$.

So that the instrument has the same directional effect on all whose behaviour it changes. Assume for instance that $D_{1i} \geq D_{0i}$; in this case the IV estimator has the very simple form:

$$\begin{aligned} \widehat{\beta}_{IV} &= \frac{cov(\ln y_i, Z_i)}{cov(S_i, Z_i)} & (18) \\ &= \frac{E[\ln y_i|Z_i = 1] - E[\ln y_i|Z_i = 0]}{E[S_i|Z_i = 1] - E[S_i|Z_i = 0]}. & (19) \end{aligned}$$

Under independence of β_i, α_i and ε_i from Z_i this reduces to

$$\begin{aligned} \widehat{\beta}_{IV} &= \frac{E[\beta_i S_i|Z_i = 1] - E[\beta_i S_i|Z_i = 0]}{\Pr[D_{1i} > D_{0i}]} & (20) \\ &= \frac{E[\beta_i D_{1i}] - E[\beta_i D_{0i}]}{\Pr[D_{1i} > D_{0i}]} & (21) \\ &= E[\beta_i|D_{1i} > D_{0i}]. & (22) \end{aligned}$$

Which provides a useful interpretation for IV - it estimates the average returns among those induced to change behaviour under the instrument - the Local Average Treatment Effect, LATE. For example, suppose $Z_i = 1$ reflected a bad financial event for the family at the time the education decision was being made. Then IV would pick out the average marginal return among those taking schooling level S_1 relative to the average returns among those taking schooling level S_1 whose family did not experience financial difficulties. This could be a very high local average return.

3.2.4 Some Drawbacks of IV

The LATE discussion highlights the point that the IV estimate in the heterogeneous returns model will typically vary depending on which instrument is used. Moreover, it could vary widely according to the local average it recovers. For example, if the instrument is a change in the minimum school leaving age which then induces a change in those achieving schooling level S_{1i}, the IV estimator will estimate the average returns among those induced to achieve $S_{1i} = 1$ by the school leaving age reform. These could be a group with very high (or very low) returns. If those who now achieve S_{1i} were those who had little to gain then the local average could be low. If on the other hand, they are individuals who had previously left education earlier because of a lack of information or family resources, the local average return for them could be quite high.

In any case the lesson to be learned from the discussion of IV in the heterogeneous returns model is that the nature of the incidence of the instrument within the distribution of returns β_i is critical in understanding the estimated β coefficient.

3.3 Control Function Methods

3.3.1 The Heterogeneous Single Treatment Model

Suppose S_{1i} in (3) is determined according to the binary response model

CF: A1 $S_{1i} = 1(Z_i'\pi + v_i > 0)$ and $v_i \sim N(0, \sigma_v^2)$

also assume α_i and β_i relate to S_{1i} according to

CF: A2 $\alpha_i - \alpha_0 = \rho_{\alpha v} v_i + \xi_{\alpha i}$

CF: A3 $\beta_i - \beta_0 = \rho_{\beta v} v_i + \xi_{\beta i}$

Note that given CF: A1 - CF: A3

$$E[(\alpha_i - \alpha_0)|S_{1i} = 1] = \rho_{\alpha v}\lambda_{1i},$$

$$E[(\alpha_i - \alpha_0)|S_{1i} = 0] = \rho_{\alpha v}\lambda_{0i}$$

and

$$E[(\beta_i - \beta_0)|S_{1i} = 1] = \rho_{\beta v}\lambda_{1i}$$

where λ_{0i} and λ_{1i} are the standard inverse Mills ratios from the normal selection model (Heckman (1979)) - or *control functions.*

With these additional assumptions in place, the model (9) can be written

$$\ln y_i = \alpha_0 + \beta_0 S_{1i} + \rho_{\alpha v}(1 - S_{1i})\lambda_{0i} + (\rho_{\alpha v} + \rho_{\beta v})S_{1i}\lambda_{1i} + \omega_i \tag{23}$$

with

$$E[\omega_i|S_{1i}, (1 - S_{1i})\lambda_{0i}, S_{1i}\lambda_{1i}] = 0.$$

Consequently, least squares estimation of the augmented log earnings regression which includes the additional terms $(1 - S_{1i})\lambda_{0i}$ and $S_{1i}\lambda_{1i}$ will produce a consistent estimator of β_0. These additional terms are labelled *control functions* and eliminate the bias induced by the endogeneity of schooling. Note that these control function terms depend on the unknown reduced form parameter π, but this can be consistently estimated at a first stage Probit step - again analogous to the selection model[8].

3.3.2 The Homogeneous Returns Model

In the special case where β_i is constant for all i, the control function terms reduce to a single term

$$\rho_{\alpha v}((1 - S_{1i})\lambda_{0i} + S_{1i}\lambda_{1i}). \tag{24}$$

3.3.3 The Multiple Treatment Model

The extension to the multiple treatment case is reasonably straightforward. As in (4), write the exhaustive set of J treatments (schooling levels) under

[8] An early example of this can be found in Willis and Rosen (1979).

examination as S_{1i}, S_{2i}, ..,S_{Ji}. The heterogeneous returns model (9) is then extended to

$$\ln y_i = \alpha_i + \sum_{j=1}^{J} \beta_{ji} S_{ji} + \sum_{j=0}^{J} \rho_j S_{ji} \lambda_{ji}.. + \beta_{Ji} S_{Ji} + \omega_i \tag{25}$$

with $S_{0i} = 1 - \sum_{j=1}^{J} S_{ji},.$ and

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

However, note that 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" 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 for *all* λ_{ji} terms.

3.3.4 The Heterogeneous One-Factor Model

Consider the one factor or "years of schooling" model (5). As mentioned above in the discussion of IV for this case, assumption IV A4 would seem plausible, so that we may write

$$S_i = Z_i' \pi + v_i \text{ where } E(v_i | Z_i) = 0. \tag{26}$$

Now given the control function assumptions **CF A1 - CF A3**, we may write

$$\ln y_i = \alpha_0 + \beta_0 S_i + \rho_{\alpha v} v_i + \rho_{\beta v} S_i v_i + \omega_i \tag{27}$$

where

$$E[\omega_i | S_i, v_i, S_i v_i] = 0$$

and now the inclusion of control functions v_i and $S_i v_i$, render least squares consistent (see Garen (1984), for example).

Again note that this is very specific to the continuous schooling measure S_i in the one factor model and the additively separable model for S_i. Finally, as we noted above, in the homogeneous one factor model the control function approach reduces to a regression of log earnings on schooling S_i including the reduced form error v_i as an additional regressor

$$\ln y_i = \alpha_0 + \beta_0 S_i + \rho_{\alpha v} v_i + \omega_i \text{ where } E[\omega_i | S_i, v_i] = 0 \tag{28}$$

The estimation of α and β by the inclusion of v_i in this homogeneous returns specification is also exactly equivalent to the IV approach.

3.3.5 Some Drawbacks of CF

The control function approach allows for heterogeneity in a multiple treatment model but at the cost of being able to construct a set of control function - one for each treatment - that have independent variation. This places strong demands on instrument availability. That is an excluded instrument is required for each treatment. Moreover, a functional form assumption is typically made on the control function. This is equivalent to making an assumption on the distribution of unobservables.

It is true that the distributional assumptions can be relaxed, following the recent developments in the semiparametric selection model literature, but the requirement on excluded instruments can only be weakened by strengthening the model for the treatment choices. For example, in the application below we exploit the sequential nature of educational qualifications to estimate an ordered Probit model from which the control functions for each qualification level can be derived. In that example a single instrument would be sufficient.

3.4 Matching Methods

The matching method is a (non-parametric) approach to the problem of identifying the treatment impact on outcomes. The main purpose of matching is to re-establish the conditions of an experiment when no such data is available. As discussed earlier, in the case of a social experiment, random assignment of individuals to treatment ensures that potential outcomes are independent of treatment status, which allows one to compare the treated and the non-treated directly, without having to impose any structure on the problem.

The matching method attempts to mimic an experiment by choosing a comparison group from all the non-treated such that the selected group is as similar as possible to the treatment group in terms of their observable characteristics. Under the matching assumption that all the outcome-relevant differences between any two individuals are captured in their observable attributes, the only remaining difference between the two groups is programme participation, so that the outcomes of the matched non-treated individuals constitute the correct sample counterpart for the missing information on the outcomes of the treated had they not been treated[9].

The central issue in the matching method is choosing the appropriate matching variables. We will point out this is a knife edge decision as there can be too many as well as too few to satisfy the assumptions for recovering a consistent estimate of the treatment effect. In some ways this mirrors the issue in choosing an appropriate excluded instrument in the IV and Control Function approaches discussed above. However, it will become clear that instruments do not make appropriate matching variables and visa versa. Instruments should satisfy an exclusion condition in the outcome equation conditional on the treatment whereas matching variables should impact on both the outcome and treatment equations.

3.4.1 General Matching Methods

To illustrate the matching solution for the average impact of the treatment on the treated in a more formal way, consider a completely general specification

[9] This discussion refers to the estimation of the average treatment effect on the treated; for the average effect of the non-treated, a symmetric procedure applies.

of the earnings outcomes in the single discrete treatment case. Denote the earnings outcome that would result if individual i were to receive the level of education of interest as $\ln y_i^1$, and let $\ln y_i^0$ be the earnings outcome if the same individual were not to receive this education level. The actually observed outcome $\ln y_i$ can thus be expressed in terms of the potential outcomes and of the observed treatment indicator S_{1i} as $\ln y_i = \ln y_i^0 + S_{1i}(\ln y_i^1 - \ln y_i^0)$. 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}:

MM: A1 $\ln y^0 \perp S_1 \mid X$

or its weaker version:

MM: A1' $E(\ln y^0 \mid X, S_1 = 1) = E(\ln y^0 \mid X, S_1 = 0)$

This assumption of *selection on observables* requires that, conditional on observed attributes X, the distribution of the (counterfactual) outcome $\ln y^0$ in the treated group is the same as the (observed) distribution of $\ln y^0$ in the non-treated group. For each treated observation $(\ln y_i : i \in \{S_{1i} = 1\})$ we can look for a non-treated (set of) observation(s) $(\ln y_i : i \in \{S_{1i} = 0\})$ with the same X-realisation. Under the matching assumption that the chosen group of matched controls - i.e. conditional on the X's used to select them - does not differ from the treatment group by any variable which is systematically linked to the non-participation outcome $\ln y^0$, this matched control group constitutes the required counterfactual. Actually, this is a process of re-building an experimental data set.

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

MM: A2 $P(S_{1i} = 1 \mid X_i) < 1$ for $X_i \in C^*$

which guarantees that all treated individuals have a counterpart on the non-treated population over the set of X values over which we seek to make a comparison. Depending on the sample 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 X does not overlap for the treated and non-treated groups, matching has in fact to be performed over the common support region; the estimated treatment effect has then to be redefined as the mean treatment effect for those treated falling within the common support.

Based on these conditions, a subset of comparable observations is formed from the original sample, and with those a consistent estimator for the treatment impact on the treated (within the common support C^*) is the empirical counterpart of:

$$\begin{aligned}
&E(\ln y^1 - \ln y^0 | S_1 = 1, C^*) \\
&= \frac{\int_{C^*}[E(\ln y^1 \mid X,\, S_1 = 1) - E(\ln y^0 \mid X,\, S_1 = 1)]\; dF(X \mid S_1 = 1)}{\int_{C^*} dF(X \mid S = 1)} \\
&= \frac{\int_{C^*}[E(\ln y^1 \mid X,\, S_1 = 1) - E(\ln y^0 \mid X,\, S_1 = 0)]\; dF(X \mid S_1 = 1)}{\int_{C^*} dF(X \mid S = 1)} \\
&= \frac{\int_{C^*}[E(\ln y \mid X,\, S_1 = 1) - E(\ln y \mid X,\, S_1 = 0)]\; dF(X \mid S_1 = 1)}{\int_{C^*} dF(X \mid S = 1)}
\end{aligned}$$

If the second assumption is fulfilled and the two populations are large enough, the common support is the entire support of both. Note that this estimator is, simply, the mean difference in earnings on the common support, appropriately weighted by the distribution of participants.

If we are also interested in using matching to recover an estimate of the treatment on the non-treated, as we do in our application to the NCDS data, we need to extend MM: A1 to include $\ln y^1$and MM: A2 to $0 < P(S_{1i} = 1 \mid X_i)$ for $X_i \in C^*$.

As it should now be clear, the matching method avoids defining a specific form for the outcome equation, decision process or either unobservable term. We simply need to ensure that, given the right observables X, the observations of non-participants are statistically what the participants' observations would have been had they not participated. Under a slightly different perspective, it might be said that we are decomposing the conditional treatment effect in the following way:

$$
\begin{aligned}
&E(\ln y^1 - \ln y^0|X, S_1 = 1) = \\
&\qquad \{E(\ln y^1|X, S_1 = 1) - E(\ln y^0|X, S_1 = 0)\} - \\
&\qquad \{E(\ln y^0|X, S_1 = 1) - E(\ln y^0|X, S_1 = 0)\}
\end{aligned}
$$

The latter term is the bias conditional on X, which under the matching assumption MM:A1' is zero.

Heckman, Ichimura, Smith and Todd (1998) use experimental data to provide a very useful breakdown of the bias which arises when the treatment on the treated parameter is estimated using the earnings of the observed group with $S_{1i} = 0$ $(\ln y^0 \mid X, S_{1i} = 0)$ to construct the counterfactual, rather than the true counterfactual $(\ln y^0 \mid X, S_{1i} = 1)$. They show that the bias term can be decomposed in three distinct parts:

$$
bias \quad = \quad E(\ln y^0 \mid X, S_{1i} = 1) - E(\ln y^0 \mid X, S_{1i} = 0) \quad = \quad B_1 + B_2 + B_3
$$

where B_1 represents the bias component due to non-overlapping support of X; B_2 is the error part due to misweighting on the common support of X as the resulting empirical distributions of treated and non-treated are not the same even when restricted to the same support; and B_3 is the true econometric selection bias resulting from "selection on unobservables". Through the process of choosing and re-weighting observations, matching corrects for the first two sources of bias. Arguing the importance of the remaining source of bias amounts to arguing the inadequacy of the conditional independence assumption (MM: A1) in the specific problem at hand, which should be done in relation to the richness of the available observables (i.e. the data X) in connection to the selection/outcome processes.

3.4.2 Propensity Score Matching

It is clear that when a wide range of variables X is in use, matching can be very difficult to implement due to the high dimensionality of the problem. A more feasible alternative based on the results of Rosenbaum and Rubin (1983) is to match on a *balancing score*, that is a function of the observables X, $b(X)$, with the property: $X \perp S_1 \mid b(X)$. This is usually carried out on the *propensity score,* the propensity to participate given the full set of observed characteristics: $p(X_i) \equiv P(S_{1i} = 1 \mid X_i)$. 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 X. Rosenbaum and Rubin have further shown that under MM: A1 and MM: A2, that is when

$$\ln y^1, \ln y^0 \perp S_1 \mid X \quad \text{and} \quad 0 < p(X) < 1$$

then

$$\ln y^1, \ln y^0 \perp S_1 \mid p(X)$$

In other words, the conditional independence remains valid if $p(X)$ - a scalar variable on the unit interval - is used for matching rather than the complete vector of X. Under the two matching assumptions, a matched sample at each propensity score $p(X)$ is thus equivalent to a random sample: conditioning on the propensity score, each individual has the same probability of assignment to treatment, as in a randomised experiment, so that individuals with the same value of $p(X)$ but a different treatment status can act as controls for each other. At any value of $p(X)$, the difference between the treatment and the non-treatment averages is thus an unbiased estimate of the average treatment effect at that value of $p(X)$, and the estimate of matching can be thought of as a weighted average of the estimates from a series of mini random experiments at the different values of $p(X)$.

3.4.3 Implementing Propensity Score Matching Estimators

The main idea of matching is to pair to each treated individual i some group of 'comparable' non-treated individuals and to then associate to the outcome $\ln y_i$ of treated i, a matched outcome $\widehat{\ln y_i}$ given by the (weighted) outcomes of his 'neighbours' in the comparison group.

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

$$\widehat{\beta}_{MM} = \sum_{i \in \{S_{1i}=1 \cap C^*\}} \left\{ \ln y_i - \widehat{\ln y_i} \right\} w_i$$

with w_i typically set equal to $1/N_1^*$ (N_1^* being the number of treated individuals falling within the common support C^*).

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

$$\widehat{\ln y_i} = \sum_{j \in C^0(p_i)} W_{ij} \ln y_j \qquad (29)$$

where

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

The different matching estimators differ in how they construct the matched outcome $\widehat{\ln y}$, that is in how they define the neighbourhood for the control group for each treated observation. They also differ in how they choose the weights for the control group.

The traditional and most intuitive form of matching is *nearest-neighbour matching,* which associates to the outcome of treated unit i a 'matched' outcome given by the outcome of the most observably similar control unit k_i. This amounts to defining $C^0(p_i)$ as a singleton:

$$C^0(p_i) = \left\{ k_i \in \{S_1 = 0\} : |p_i - p_{k_i}| = \min_{j \in \{S_1=0\}} \{|p_i - p_j|\} \right\}$$

and setting $W_{ij} = 1(j = k_i)$ (ie. giving a unity weight to the closest control observation and zero to any other).

In our application below we use a variant of nearest-neighbour matching, *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 allowing to better enforce common support on the propensity score. This involves matching treated individual i with its nearest-neighbour non-treated individual j provided that:

$$\delta > |p_i - p_j| = \min_{k \in \{S_1=0\}} \{|p_i - p_k|\}$$

If none of the non-treated individuals are within a certain predefined absolute distance or caliper δ of the treated individual i under consideration, individual i is left unmatched.

A different class of matching estimators has been recently proposed by Heckman, Ichimura and Todd (1997) and (1998), Heckman, Ichimura, Smith and Todd (1998). In *kernel-based matching*, the outcome $\ln y_i$ of treated individual i is matched to a weighted average of the outcomes of more (possibly all) non-treated individuals, where the weight given to non-treated individual j is in proportion to the closeness of the propensity scores of i and j. That is, the weight in equation (29) above is set to:

$$W_{ij} = \frac{K\left(\frac{p_i - p_j}{h}\right)}{\sum_{j \in C^0(p_i)} K\left(\frac{p_i - p_j}{h}\right)}$$

where h is the bandwidth and $K(.)$ is the kernel.

With e.g. the Gaussian kernel, $K(u) \propto \exp\{-u^2/2\}$ and all the non-treated units are used to smooth at p_i, that is $C^0(p_i) = \{j : S_{1j} = 0\}$. By contrast, with the Epanechnikov kernel, $K(u) \propto (1 - u^2) \cdot 1(|u| < 1)$ and thus only those non-treated units whose propensity score falls within a fixed 'caliper' of h from p_i are used to smooth at p_i, that is $C^0(p_i) = \{j \in \{S_1 = 0\} : |p_i - p_j| < h\}$.

Typically with kernel-based matching 'common support' is imposed on treated individuals, that is those treated whose propensity score is larger than the largest propensity score in the non-treated pool are left unmatched.

3.4.4 Multiple Treatments - Mahalanobis metric matching

The most attractive feature of propensity score matching methods is the fact that they allow to select individuals based on a single variable (nearest-neighbour matching), or to non-parametrically smooth outcomes on a single variable (kernel-based matching). The idea underlying *Mahalanobis metric matching* is to reduce the dimensionality of the matching problem by first combining the variables one wants to match on into a distance measure and to then match on the resulting scalar variable. Choosing the Mahalanobis distance defined below has the advantage of offering a unit free metric, which is essential when the matching variables have different units. Rubin (1979) and (1980) and Rosenbaum and Rubin (1985) have looked at Mahalanobis metric matching as an alternative to, as well as in combination with, propensity score matching, by comparing the performance of matching estimators based on the propensity score alone, on the X's combined into a distance measure or on the X's together with the propensity score combined into a distance measure.

Even when one decides in favour of propensity score matching, however, in some settings one may need to match on more than one propensity score. In particular, when individuals can receive a set of different treatments, adequately controlling for observed differences may require matching on two scores (for a very recent application to a framework of mutually exclusive multiple treatments, see Lechner (2001)). In our application to the multiple treatment framework below, we will need to match on three scores.

In such situations, the different scores can be first combined into a distance metric. Formally, if matching needs to be performed on $k > 1$ scores, the formulae for the different types of estimators described above continue to apply after replacing $p_i - p_j$ with the Mahalanobis distance $d(i, j)$ defined as:

$$d(i, j) = (\mathbf{P}_i - \mathbf{P}_j)'\mathbf{W}^{-1}(\mathbf{P}_i - \mathbf{P}_j)$$

where $\mathbf{P}_i$ is the $k \times 1$ vector of scores of individual i, $\mathbf{P}_j$ is the $k \times 1$ vector of scores of individual j and $\mathbf{W}$ is the pooled within-sample $(k \times k)$ covariance matrix of $\mathbf{P}$ based on the sub-samples of the treated and complete non-treated pool. This metric effectively weights each co-ordinate in proportion to the inverse of the variance of that co-ordinate.

3.4.5 Some Drawbacks to Matching

The most obvious criticism that may be directed to the matching approach is the fact that the identifying conditional independence assumption (MM: A1) on which the method relies is in general a very strong one. As mentioned above, the plausibility of such an assumption should always be discussed on a case-by-case basis, with account being taken of the informational richness of the available dataset (X) in relation to the institutional set-up where selection into the treatment takes place.

Furthermore, the common support requirement implicit in MM: A2 may at times prove quite restrictive. In the case of social experiments, randomisation generates a comparison group for each X in the population of the treated, so that the average effect of the treatment can be estimated over the entire support. By contrast, under the conditional independence assumption matching generates a comparison group, but only for those X values that satisfy MM: A2. In some cases, matching may not succeed in finding a non-treated observation with similar propensity score for all of the participants. If MM: A2 fails for some subgroup(s) of the participants, the estimated treatment effect has then to be redefined as the mean treatment effect for those treated falling within the common support.

If the impact of treatment is homogeneous, at least within the treated group, no additional problems arises besides the loss of information. Note, however, that the setting is general enough to include the heterogeneous case. If the impact of participation differs across treated individuals, restricting to the common subset may actually change the parameter being estimated; in other words, it is possible that the estimated impact does not represent the mean outcome of the programme, so that we are unable to identify β_T. This is certainly a drawback of matching in respect to randomised experiments; when compared to standard parametric methods, though, it can be viewed as the price to pay for not resorting to the specification of a functional relationship allowing to extrapolate outside the common support. In fact, the absence of good overlap may in general cast doubts as to the robustness of traditional methods relying on functional form.

4 Measuring Returns in the NCDS: The Data

In this section we compare the different estimation approaches outlined in Section 3. We begin by comparing the wage outcomes of those individuals who obtain some qualifications (qualifications) with those who obtain no qualifications (no qualifications). We do this so that we can compare different regression (parametric) based estimates of the returns to obtaining qualifications with *sin-*

gle treatment matching techniques. We then go on to consider the sequence of multiple treatments (O levels or equivalent, A levels or equivalent, higher education) and again compare parametric based estimates of the returns to these different education paths, with estimates derived from multiple treatment matching approaches.

4.1 Single Treatment Models - qualifications versus no qualifications

The estimated returns to obtaining qualifications versus leaving education with no qualifications are shown in Table 1. The "treated" are all those obtaining at least an O-level qualifications or equivalent and we would expect a wide range of returns among this group. We begin by comparing the different regression based estimates. Specification (i) gives the OLS estimate when we only use minimal controls (region and ethnicity). We see from Table 1 that for this basic specification the estimated return to staying on for men is 37.0%. When we include ability and school type variables (specification (ii)) these estimates fall. The estimate for is reduced to 27.5%. There is a further fall when we also include standard family background variables[10], with the estimated return for this sample of men being 25.5%.

The Instrumental Variable and Control Function Results: The next two specifications in Table 1 allow the education variable to be endogenous. In the first of these, specification (iv), we include controls for region and ethnicity alone (i.e. the same as specification (i)) and use the standard family background variables (set I) as instruments for staying-on. The IV estimates of the returns to staying on are significantly higher at 78.6%. However, we should note that the overidentifying restrictions are overwhelmingly rejected suggesting misspecifica-

[10] The family background variables that we include are parent's education, age, education×age, father's social class when child was 16 (six dummies), mother's employment status when the child was 16, and the number of siblings the child had at 16.

tion. In specification (v) we instead include these family background variables as controls in the regression together with ability and school type variables(as in specification (iii)) and use more credible instruments, set II. These are the number of older siblings the child has (which controlling for total number of siblings is exogenous), whether the family was experiencing financial difficulties in 1969 or 1974, and finally parental interest in the child's education at the age of seven as assessed by the child's teacher[11]. The IV estimates using this set II instruments are still higher than least squares with controls but have fallen back to 47.4%[12].

The large increase in estimated returns over the corresponding OLS specifications probably reflect three factors. These factors are always important to consider when interpreting IV estimates of returns. The first relates to the discussion of the Local Average Treatment Effect - LATE. In that discussion we noted that the IV estimator, in the heterogeneous returns single treatment model, will depend on the choice of instrument and the group of individuals whose schooling is impacted by the instrument. Here the important instruments for specification (iv) relate to parent's education and social class, whereas in specification (v) they relate to birth order, parent's interest in the child's education and bad financial problems for the family. It could well be that the average return within this group who would otherwise have completed qualifications is high. The third relates to measurement error. This would also result in an increase over the OLS estimates but it is unlikely that there is serious measurement error in these recorded education levels. The third factor relates to the use of invalid instruments. This will induce an upward bias in the return. Some experiments we have conducted suggest that certain of the instruments used in specification (iv)- fathers and mother's education, for example - may be

[11] For both the mother and father we have 3 dummy variables corresponding to "expects too much", "very interested" and "some interest".

[12] The Sargan test does not reject the validity of the instruments: p-value 0.317.

invalid instruments.

Overall the high values for these IV estimates probably reflect the extreme heterogeneous nature of the group who obtain qualifications and emphasise the extreme caution with which instrumental variable estimates of rates of return should be treated. This is confirmed by the results from the control function model which specifically allows for heterogeneity in the returns to qualifications. This control function model is simply a more general version of the endogenous treatment effects model, in that the only difference is that the inverse mill ratio is now fully interacted with our qualifications dummy. Our estimate of the return to qualifications is further reduced to 37.7%. We reject the null hypothesis of homogeneous returns for those with qualifications[13]. These estimates are almost identical to those of specification (i) where we have only controlled for region and ethnicity. In the full multiple treatment model below, the corresponding IV estimates remain higher than the OLS results, however they appear more reasonable and stable across specifications.

The Matching Results: We now move on to our matching estimates. We adopt two sorts of matching, one-to-one matching with replacement (nearest neighbour matching) and kernel based matching using the Epanechnikov kernel. We estimate a probit model of qualifications versus no qualifications and match on the predicted probability of obtaining qualifications from this model. We estimate the effect of obtaining qualifications for those who actually obtained these qualifications (treatment on the treated) and for those who did not obtain qualifications (treatment on the non-treated). The first involves matching each person who obtained a qualification, with somebody who looks like them, but did not obtain any qualifications. The differences in wages between these two groups (each of equal size) is the estimated effect of obtaining a qualification

[13] The coefficients on the control functions (see section 3) are -0.003 (0.047) for no qualifications and -0.278 (0.109) for qualifications.

for those who did so (those who did not obtain a qualification can be used more than once). With kernel density matching, the wage of each individual who obtained a qualification is compared with the weighted average of wages for all individuals who did obtain a qualification who fall within a specified bandwidth. Again the difference in wage outcomes gives us our estimate of the effect of obtaining a qualification, for those who did undertake a qualification. The second involves matching everyone who did not obtain a qualification, with a person who looked like them, but did obtain a qualification. The difference between the outcomes for these two groups gives an estimate of what this group would have received, if they had instead decided to undertake a qualification. To ensure common support, we impose a caliper of 0.0025 for our nearest neighbour matching and a bandwidth of 0.06 for our kernel based matching[14]. We also report the estimate when common support is not imposed.

If we look at the results for men in Table 1 we see that the estimated return to obtaining a qualification for those who actually undertook the qualification (treatment on the treated) is around 31 to 35 %[15], whereas the estimated return for those who did not (treatment on the non-treated) is around 23 to 24 %, a significant difference of around 10 percentage points. The treatment on the treated results are somewhere between our highest and lowest OLS estimates but below our IV and CF estimates whereas the treated on the non-treated results are below all these estimates. These results suggest that if those who have no qualifications were to undertake them, they would receive a lower return than the group who had already undertaken them.

[14] We experimented with the sensitivity of our estimates to the caliper and bandwidths used but this made little difference to our estimates.

[15] We focus on the results where common support has been imposed.

Table 1: The returns to obtaining a school qualification - NCDS Men

Estimation	Coef	(S.E.)	No.
technique and specification:			obs
OLS:			
(i) region and ethnicity	0.370	(0.017)	3639
(ii) – (i) + ability & school type variables	0.275	(0.017)	3639
(iii) – (ii) + family background (excl instruments)	0.255	(0.017)	3639
Instrumental Variables:			
(iv) – (i) using set I instruments	0.786	(0.038)	3639
(v) – (iii) using set II instruments	0.474	(0.082)	3639
Control Function:			
(vi) – (iii) using set II instruments	0.377	(0.094)	3639
Matching - *Effect of treatment on the treated:*			
(vii) Nearest neighbour - all obs	0.346	(0.034)	2988
(viii) Nearest neighbour - caliper (0.0025)	0.352	(0.033)	2815
(ix) Kernel weighting - all obs	0.334	(0.025)	2988
(x) Kernel weighting - bandwidth(0.06)	0.313	(0.025)	2795
Matching - *Effect of treatment on the non-treated:*			
(xi) Nearest neighbour - all obs	0.231	(0.028)	651
(xii) Nearest neighbour - caliper(0.0025)	0.242	(0.028)	619
(xiii) Kernel weighting - all obs	0.235	(0.020)	646
(xiv) Kernel weighting -bandwidth (0.06)	0.236	(0.019)	643

Note: Matching standard errors are bootstrapped (390 replications)

4.2 The Multiple Treatment Model

We now turn to a more disaggregated analysis that focuses on the sequential nature of educational qualifications. To this end we separate the qualifications variable into those who stopped education after completing O levels or equivalent, those who stopped after completing A levels or equivalent, and those who completed O levels, A levels *and* higher education (HE). In this sequential model we exploit the ordering of our outcomes. For the control function estimates we use an ordered probit model in our first stage for these three levels of education over the base[16]. Our matching estimator uses an adaption of the propensity score matching method for multiple sequential treatments and this is described in detail in Appendix A.

The Instrumental Variable and Control Function Results: The OLS, IV and control function results are shown in Table 2. The return to higher education versus no qualifications is obtained by adding the return to O levels or equivalent, the return to an A level or equivalent and the return to HE. Overall the results have a distinct pattern. Controlling for ability and school type is important and reduces the return to education at all levels. The IV results are higher than the OLS results. The Control Function estimates are somewhat smaller than the IV estimates but lie above the OLS estimates.

These results are enormously informative. The fact that OLS estimates of the impact of education on earnings are lower than either the IV or CF estimates suggests that there is some attenuation bias in OLS most likely due to measurement error. The finding that IV is often higher than CF is strong evidence of heterogeneous responses, the IV estimate picking out some local average effect

[16] From this we calculate the control function terms described in Section 3. In this approach the control functions are interacted with each of the qualification levels. Standard errors are corrected to take into account the generated regressor(s) in the model as well as for heterescedasticity.

Table 2: Regression estimates of the returns to qualifications - NCDS Men

Variable	(i) OLS Basic Specification	(ii) OLS Full set of controls	(iii) IV Spec (i)	(iv) IV Spec (ii)	(v) CF Spec (ii)
O level or equivalent	0.207	0.166	0.315	0.289	0.221
	(0.018)	(0.018)	(0.053)	(0.010)	(0.010)
A level or equivalent	0.094	0.077	0.170	0.166	0.155
	(0.016)	(0.016)	(0.035)	(0.068)	(0.137)
HE	0.298	0.249	0.393	0.373	0.331
	(0.016)	(0.016)	(0.029)	(0.063)	(0.153)
Number of observation	3639	3639	3639	3639	3639

which, as we have already indicated, can be quite high in certain subpopulations. Among these estimates the CF estimate is probably the most reliable estimate of the average impact of each education level on gross earnings[17]. The results show significant overall returns to educational qualifications at each stage of the educational process even after correcting for detailed background variables and ability differences, as well as allowing for heterogeneity in the education response parameters.

The Matching Results: To implement our adapted propensity score matching method for this sequential treatment case, we estimated three propensity scores. We used the full set of variables available for matching but do not include our set II instruments which we argued impacted on the education treatment level but not directly (conditional on treatment) on the earning outcome. Such instruments are not appropriate matching variables. The first was a probit of qualifications (O levels, A levels and HE) versus no qualifications; the second was a probit of A-levels and HE versus O levels; and the third was a probit of higher education versus A levels. On the basis of these three probits we estimated the unconditional probability of undertaking each of our 3 treatments and no qualifications (which in Appendix A we denote by $r(j, X)$, $j - 0, 1, 2, 3$) for *our whole sample,* including those who did not make particular qualification

[17] The coefficients on the control functions are -0.061 (0.070) for no qualifications; -0.099 (0.056) for O levels; -0.121(0.063) for A levels; and -0.084 (0.067) for HE.

transitions. We then compared the outcomes across each of our four groups, matching on the appropriate *one* dimensional propensity score for the particular transition in question which again is defined in Appendix A.

Our approach involves estimating the incremental return to each of our three qualifications by actual qualification. For those with no qualifications, we estimate the returns they would have got if they had undertaken each of the three qualifications (treatment on the non-treated). For those with O level qualifications we estimate the return they obtained for taking that qualification (treatment on the treated) and the returns they would have obtained if they had progressed to A levels or HE (treatment on the non-treated). For those with A levels we estimate the returns they obtained for undertaking O and A level qualifications (treatment on the treated) and the returns they would have obtained if they had progressed to HE (treatment on the non-treated). For those with HE all estimates are treatment on the treated. In all of our tables we italicise treatment on the non-treated results.

We present three sets of results. The first set of results uses nearest neighbour matching but does not impose common support. The second set uses nearest neighbour matching but only accepts matches within a caliper of 0.0025 *and* only includes individuals who are matched for every possible transition (so that we can make comparisons across the same sets of individuals). The third set uses kernel density matching using a bandwidth of 0.06 and again only includes individuals who are matched at every possible transition. The results of doing this are given in Table 3.

Focusing on the results in Table 3, we see that the estimate of a return to an O level is generally lower when we impose common support (with the exception of the effect estimated using nearest neighbour for those with no qualifications). If we focus on the treatment on the treated results the estimates of the returns are very close using both methods and rise with the actual qualification

Table 3: Multiple Treatment Matching Results - NCDS Men

Actual	Estimated Returns						Number
Qualification:	O level		A level		HE		of
	Est	(SE)	Est	(SE)	Est	(SE)	observations
None:							
Nearest neighbour - all obs	*0.245*	*(0.032)*	*0.023*	*(0.042)*	*0.286*	*(0.052)*	651
Nearest neighbour - cal (0.0025)	*0.270*	*(0.038)*	*0.054*	*(0.045)*	*0.258*	*(0.047)*	443
Kernel denisty - bdwith (0.06)	*0.150*	*(0.021)*	*0.065*	*(0.023)*	*0.273*	*(0.035)*	619
O level:							
Nearest neighbour - all obs	0.189	(0.030)	*0.179*	*(0.025)*	*0.226*	*(0.034)*	993
Nearest neighbour - cal (0.0025)	0.177	(0.032)	*0.185*	*(0.031)*	*0.235*	*(0.036)*	716
Kernel denisty - bdwith (0.06)	0.174	(0.022)	*0.066*	*(0.018)*	*0.257*	*(0.021)*	970
A level:							
Nearest neighbour - all obs	0.231	(0.039)	0.062	(0.026)	*0.367*	*(0.028)*	965
Nearest neighbour - cal (0.0025)	0.208	(0.039)	0.053	(0.031)	*0.368*	*(0.033)*	726
Kernel denisty - bdwith (0.06)	0.187	(0.026)	0.064	(0.018)	*0.260*	*(0.020)*	922
HE:							
Nearest neighbour - all obs	0.267	(0.064)	0.076	(0.045)	0.240	(0.031)	1030
Nearest neighbour - cal (0.0025)	0.250	(0.049)	0.036	(0.045)	0.263	(0.036)	595
Kernel denisty - bdwith (0.06)	0.241	(0.041)	0.077	(0.024)	0.227	(0.021)	897

Note: Treatment on the non-treated estimates italicised. Matching standard errors are bootstrapped (390 replications) obtained. The estimates range from 17.4% for those with O levels to 25% for those with HE qualifications. They are very close (on average) to the results we obtained from the control function method (22.1%). The results for treatment on the non-treated are more ambiguous. The nearest neighbour method estimates an effect of 27.0% compared to an estimate of 15.0% using the kernel density method. This illustrates clearly the problems in choosing an appropriate matching strategy and the sensitivity of some matching results to the estimation technique chosen. In our results it appears to be a particular problem with some of the treatment on the non-treated estimates.

If we now turn to the estimates of the returns to A levels we see that the treatment on the treated returns range between 3.6% and 7.7% which are unambiguously lower than all our regression based estimates. Most of the treatment on the non-treated estimates range between 5.4% to 6.6%, but there is an estimate of 18.5% for those with O levels using nearest neighbour techniques. Finally if we look at the return to HE estimates we see that with the exception of nearest neighbour matching for those with A levels (which has an estimate of 36.8%), the estimates range between 22.7% and 27.3%. These latter estimates are close to the OLS estimates whereas the former result is closer to the IV and CF results. Interestingly, for HE the matching results suggest the effect of treatment on the non-treated may be higher than treatment on the treated which has very interesting policy implications. This dimension was missed in our single treatment model.

5 Summary 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. Also the importance of allowing for heterogeneous returns - that is returns that vary across individuals for the same educational qualification. We considered three main estimation methods which rely on different identifying assumptions - instrumental variable methods, control function methods and propensity score matching methods. In each case the properties were analysed distinguishing between a single treatment model and a model where there are a sequence of possible treatments. The sequential multiple treatment model we argued is well suited to the education returns formulation where educational qualification levels in formal schooling tend to be cumulative.

With heterogeneous returns defining the 'parameter of interest' is central. We distinguished four possible parameters of interest: the treatment on the treated, the local average treatment effect, the average treatment effect and the impact of treatment on the non-treated. In the homogeneous effects model these were all equal but in the heterogeneous effects model they can differ substantially. Which is of interest will depend on the policy question. Moreover, different estimation methods were shown to identify different parameters of interest.

Our application was to NCDS 1958 birth cohort study for Britain. This data is ideally suited for evaluating the impact of education on earnings using non-experimental data and is sufficiently rich to allow the comparison of matching, control function and instrumental variable methods. First there are extensive and commonly administered ability tests at early ages. Second there are accurately measured family background and school type variables ideal for matching. Finally there are variables that are very likely to influence schooling but not wage outcomes, such us temporary financial difficulties among the parents and the composition of siblings. These make good choices for excluded instruments in the application of instrumental variables or control function methods.

The application showed the importance of allowing for schooling choices to depend on family background variables, financial constraints, etc. - as well as ability. Among the estimators used we found that OLS was very sensitive to the inclusion of ability measures and family background variables. In general these tended to reduce the estimated return. But there was strong evidence of heterogeneous returns with the instrumental variable estimator often much higher than OLS and by more than could be explained from attenuation bias caused by measurement error in the qualification level data.

References

[1] Altonji, J. and T. Dunn (1996), "Using Siblings to Estimate the Effect of Schooling Quality on Wages", *Review of Economics and Statistics*, 78, 665-671.

[2] Angrist, J. D. and Krueger, A. B. (1991), "Does Compulsory Schooling Attendance Affect Schooling Decisions", *Quarterly Journal of Economics*, 106(4), 970-1014.

[3] Angrist, J. D. and Krueger, A. B. (1992), "Estimating the payoff to schooling using the Vietnam-era draft lottery", National Bureau of Economic Research, Working Paper no. 4067.

[4] Ashenfelter, O. and Zimmerman, D. (1993), "Estimates of the returns to schooling from sibling data: fathers, sons and brothers", Princeton University, Industrial Relations Section, Working Paper no. 318.

[5] Behrman, J.R., Rosenzweig, M.R. and P. Taubman (1996), "College Choice and Wages: Estimates using Data on Female Twins", *Review of Economics and Statistics*, 78, 665-671.

[6] Blundell, R., Dearden, L., Goodman, A. and Reed, H. (2000), "The Returns to Higher Education in Britain: Evidence from a British Cohort", *Economic Journal*, 110, F82-F99.

[7] Bonjour, D., L. Cherkas, J. Haskel, D. Hawkes and T Spector (2000), "Estimating Returns to Education using a New Sample of UK Twins", mimeo, QMW, June.

[8] Bound, J., D. Jaeger and R. Baker, (1995), "Problems with Instrumental Variables Estimation when the Correlation between the Instruments and the Endogenous Explanatory Variable is Weak", *Journal of the American Statistical Association*, 90, June, 443-50.

[9] Butcher, K. and Case, A. (1994), 'The effect of sibling sex composition on women's education and earnings', *Quarterly Journal of Economics*, vol. 109, pp. 531-63.

[10] Card, D. (1999),"The Causal Effect of Education on Earnings", in O. Ashenfelter and D. Card, Handbook of Labor Economics, Vol 3, Elsevier-North Holland.

[11] Cochran, W. and D.B.Rubin (1973), 'Controlling Bias in Observational Studies', *Sankyha*, 35, 417-446.

[12] Dearden, L. (1999a), "The effects of families and ability on men's education and earnings in Britain", *Labour Economics*, vol. 6, pp. 551-567.

[13] Dearden, L. (1999b), "Qualifications and Earnings in Britain: How reliable are conventional OLS estimates of the returns to education?", IFS Working Paper No. 99/7.

[14] Dehejia, R.H. and Wahba, S. (1999), 'Causal Effects in Non-Experimental Studies: Re-Evaluating the Evaluation of Training Programmes', *Journal of the American Statistical Association*, 94, 1053-1062.

[15] Garen, J. (1984),"The returns to Schooling: A Selectivity Bias Approach with a Continuous Choice Variable", *Econometrica*, 52(5), 1199-1218.

[16] Griliches, Z. (1977), "Estimating the returns to schooling: some econometric problems", *Econometrica*, vol. 45, pp. 1-22.

[17] Harmon, C. and Walker, I. (1995), "Estimates of the economic return to schooling for the UK", *American Economic Review*, 85, 1278-86.

[18] Heckman, J. J. (1979), 'Sample selection bias as a specification error', *Econometrica*, vol. 47, pp. 153-61.

[19] Heckman, J.J., Ichimura, H. and Todd, P. (1997), 'Matching As An Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme', *Review of Economic Studies*, 64, 605-654.

[20] Heckman, J.J., Ichimura, H., and Todd, P (1998), 'Matching as an Econometric Evaluation Estimator', *Review of Economic Studies*, 65, 261-294.

[21] Heckman, J.J., Ichimura, H., Smith, J. and Todd, P. (1998), 'Characterizing selection bias using experimental data', *Econometrica*, vol. 66, pp. 1017-98.

[22] Heckman, J.J., R. Lalonde and J. Smith (1999), "The Economics and Econometrics of Active Labor Market programs" in O. Ashenfelter and D. Card, Handbook of Labor Economics, Vol 3, Elsevier-North Holland.

[23] Heckman, J. And Robb, R. (1985), "Alternative methods for Evaluating the Impact of Interventions", in Longitudinal Analysis of Labour market Data (New York: Wiley).

[24] Heckman, J., Smith, J. And N. Clements, (1997), "Making the Most out of Program Evaluations and Social Experiments: Accounting for Heterogeneity in program Impacts", *Review of Economic Studies*, 64, 487-536.

[25] Imbens, G. (2000), "The Role of Propensity Score in Estimating Dose-Response Functions", *Biometrika*, 87, 3, 706-710.

[26] Imbens, G. and Angrist, J. (1994), "Identification and Estimation of Local Average Treatment Effects", *Econometrica*, 62, 2, 467-476.

[27] Lechner, M. and Miquel, R. (2001), 'A Potential Outcome Approach to Dynamic Programme Evaluation - Part I: Identification', discussion paper, SIAW, Universität St. Gallen.

[28] Lechner, M. (2001), "Identification and Estimation of Causal Effects of Multiple Treatments under the Conditional Independence Assumption, in: Lechner, M., Pfeiffer, F. (eds), *Econometric Evaluation of Labour Market Policies,* Heidelberg: Physica/Springer, p. 43-58.

[29] Meghir, C. and M. Palme (2000), “Estimating the Effect of Schooling on Earnings Using a Social Experiment”, IFS Working Paper 99/12, March.

[30] 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.

[31] Rosenbaum, P.R. and Rubin, D.B. (1985), ‘Constructing a Comparison Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score’, *The American Statistician*, 39, 33-38.

[32] Rubin, D.B. (1979), ”Using Multivariate Matched Sampling and regression Adjustment to Control Bias in Observational Studies”, *Journal of the American Statistical Association*, 74, 318-329.

[33] Rubin, D.B. (1980), ‘Bias Reduction Using Mahalanobis-Metric Matching’, *Biometrics*, 36, 293-298.

[34] Staiger, D. and J. H. Stock (1997), “Instrumental Variables Regressions with Weak Instruments”, *Econometrica*, 65(3), May, 557-86.

[35] Smith, R, and Blundell, R.W. (1986), “An exogeneity Test for a Simultaneous Tobit Model with an Application to Labour Supply”, *Econometrica*, 54, 679–685.

[36] Willis, R. J. and Rosen, S. (1979), “Education and self-selection”, *Journal of Political Economy*, vol. 87, pp. S1-S36.

A Appendix A: A Balancing Score for Sequential Multiple Treatments

In the educational context, we can view the sequential treatments (basic education, O-levels, A-levels, higher education) in a dose-response framework (cf. Imbens, 2000). Like a drug which can be applied in different doses, the sequential treatments would thus correspond to ordered levels of a treatment – education (or investment in human capital). We focus on continuous education, where individuals take uninterrupted sequential decisions of an incremental nature: at each point, they can either stop or move on to the next educational level.

We consider four treatments: $D \in \{0, 1, 2, 3\}$

- $D = 0$ for stopping at basic (i.e. no qualifications)
- $D = 1$ for stopping at O-Levels (i.e. stay on and stop with O-levels)
- $D = 2$ for stopping at A-Levels (i.e. stay on, take O-levels and stop with A-levels)
- $D = 3$ for stopping at Higher Education (i.e. stay on, take O-levels, take A-levels and stop with HE)

Incidentally, we can link this analysis to the dynamic programme evaluation framework recently suggested by Lechner and Miquel (2001). In our case of an obliged chain of educational choices, we only have a restricted set of possible sequences, four in fact. In addition, in each period there is only one type (and in fact a different type) of treatment available.

Consider four periods: in period 0 everyone achieves basic qualifications, in period 1 the relevant choice is whether to take O-levels or not; in period 2 the only treatment available is A-levels but only provided one has achieved O-levels in the previous period; while the treatment in period 3 is higher education but available only for those with A-levels. Outcome Y is then observed after period 3 (at age 33). The four possible sequences, corresponding to the four values of D defined above, are thus:

t	**0**	**1**	**2**	**3**
$D = 0$	1	0	0	0
$D = 1$	1	1	0	0
$D = 2$	1	1	1	0
$D = 3$	1	1	1	1

Define the sequence of treatments, each received at the beginning of each period t, as $S = (S_0, S_1, S_2, S_3)$, in our case, $S_0 = 1$ for all; for $t = 1, 2, 3$, $S_t \in \{0, 1\}$ and $S_t = 1 \Longrightarrow S_{t-1} = 1$.

The sequential nature of the decision process is captured by the modelling of the choice probabilities as follows:

Define

- $r(j,x) \equiv \Pr\{D=j|X=x\}$ for $j=0,1,2,3$.

Further define the following probabilities:

- $P^1(x) \equiv \Pr\{(OL \vee AL \vee HE)=1|X=x\}$ - the stay-on probability
- $P^2(x) \equiv \Pr\{(AL \vee HE)=1|OL=1,X=x\}$
- $P^3(x) \equiv \Pr\{HE=1|AL=1,X=x\}$

Thus we have:

$$D = \begin{cases} 0 & r(0,x) = 1 - P^1(x) \\ 1 & r(1,x) = (1-P^2(x))P^1(x) \\ 2 & r(2,x) = (1-P^3(x))P^2(x)P^1(x) \\ 3 & r(3,x) = P^3(x)P^2(x)P^1(x) \end{cases}$$

Let Y^k denote the outcome, $Y \equiv \ln y$, if the individual were to receive treatment (or education level) k.

We are interested in the following 12 pairwise comparisons of the effects of treatments (education levels) m and l, with $m,l \in \{0,1,2,3\}$ and $m>l$:

$$E(Y^m - Y^l \mid D=j)$$

for $j=m$ (effect on the treated) and $j=l$ (effect on the non-treated).

Note that in the framework by Lechner and Miquel (2001), evaluating the impact of stopping at one education level versus stopping at another one for those individuals who have stopped at one of such levels amounts to evaluating the effect of one sequence of treatments compared to another sequence of treatments (each of length four) for those individuals who have followed a given sequence.

In our case, we need to identify all the counterfactuals of the type:
$E(Y^m \mid D=l)$ and $E(Y^l \mid D=m)$
(More precisely, we need to identify $E(Y^k \mid D=j)$ for
$(k,j)=(0,1),(0,2),(0,3),(1,2),(1,3),(2,3)$ (effect on the treated)
as well as for
$(k,j)=(1,0),(2,0),(3,0),(2,1),(3,1),(3,2)$ (effect on the non-treated).)

An extension of the conditional independence assumption MM: A1 that would allow to identify them is what Imbens (2000) termed 'strong unconfoundedness':

$$\{Y^0,Y^1,Y^2,Y^3\} \perp D|\ X_0 = x_0$$

This CIA corresponds to Lechner and Miquel (2001) full CIA Assumption 2-I or 2-II, and can be rewritten in terms of potential outcomes each one corresponding to one of the sequences of treatments:

$$Y^{1000}, Y^{1100}, Y^{1110}, Y^{1111} \perp S_t \mid X_0 \quad \text{for all } t = 1, 2, 3$$

In words, conditional on the information observed prior to period 0, X_0, assignment to treatment in each period is independent of potential outcomes, in particular it is not affected by any new information related to the outcomes that may arrive in between schooling choices.

This implies that the complete treatment sequence, in our case the maximum level of education attained, is chosen at the beginning of period 0 – just as the dose of a drug is decided at the start – based on the information contained in X_0.

The assumption that subsequent schooling choices are not affected by the outcomes of the schooling decisions in the previous periods hinges on the absence of intermediate outcomes on which to possibly base future S decisions. This amounts to ruling out 'intermittent' educational choices - where an individual achieves a level of education, drops out of the education system, observes the corresponding outcomes (both in terms of Y and of possibly endogenous X's) and then possibly decides on re-entering the schooling system for investment in the next level of education.

Note however that the weaker form (implied by strong unconfoundedness) would suffice to our purposes:

$$\begin{gathered}\{Y^l, Y^m\} \perp D| \ X_0 = x_0, D \in \{l, m\} \\ \text{for} \ \ (l, m) = (0, 1), (0, 2), (0, 3), (1, 2), (1, 3), (2, 3)\end{gathered}$$

This relaxes the CIA above by requiring conditional independence to hold only for the subpopulations receiving treatment m or treatment l (see Lechner 2001).

The common support assumption corresponding to MM: A2 is:

$$0 < r(j, x) \equiv \Pr\{D = j | X_0 = x\} < 1 \ \text{ for } x \in C^* \text{ and } \ j = 0, 1, 2, 3$$

which for the $P^{j\prime}s$ implies the requirements:

$$0 < P^j(x) < 1 \text{ for } x \in C^* \text{ and } \ j = 1, 2, 3.$$

A.1 Looking for a balancing score

If we wanted the $X's$ (in what follows we drop the time-0 subscript from X) to be simultaneously balanced *in the four groups* defined by the highest level of education attained, i.e. if we required the same distribution in four selected (matched) subgroups of the four types of treated, we would need to look for a function of the $X's$, $b(X)$, such that (cf. Theorem 2 by Rosenbaum and Rubin, 1983 and Proposition 1 in Lechner 2001):

$$\begin{aligned} & X \perp D \mid b(X) \\ \iff \quad & E(\Pr\{D=m|X\} \mid b(X)) = \Pr\{D=m|X\} \quad \forall m = 0,1,2,3 \end{aligned}$$

Setting up the corresponding system:

$$\begin{cases} E(\Pr\{D=0|X\} \mid b(X)) = & r(0,x) = & 1-P^1(X) \\ E(\Pr\{D=1|X\} \mid b(X)) = & r(1,x) = & (1-P^2(X))P^1(X) \\ E(\Pr\{D=2|X\} \mid b(X)) = & r(2,x) = & (1-P^3(X))P^2(X)P^1(X) \\ E(\Pr\{D=3|X\} \mid b(X)) = & r(3,x) = & P^3(X)P^2(X)P^1(X) \end{cases}$$

Choosing either $b(X) = \{r(1,x), r(2,x), r(3,x)\}$ or $b(X) = \{P^1(X), P^2(X), P^3(X)\}$ would solve the system. (Note that the dimensionality has been reduced by one; this is allowed by the adding up of the treatment probabilities).

We are however just interested in the *pairwise* comparison of the various levels of the treatment, so that the above balancing score may actually be more restrictive than required for some type of comparison.

A.1.1 A balancing score for the pairwise comparisons

1. $E(Y^1 - Y^0 \mid D=j)$ for $j = 0,1$

   In this case, we just need

   $$X \perp D \mid b(X),\ D \in \{0,1\}$$

   which is verified if

   $$E(\Pr\{D=1|X, D \in \{0,1\}\}| \, b(X)) = \Pr\{D=1|X, D \in \{0,1\}\}$$

   One could use the propensity score:

   $$\Pr\{D=1|X,\ D \in \{0,1\}\} = \frac{r(1,X)}{r(1,X)+r(0,X)}$$

which is itself a function of $P^1(X)$ and $P^2(X)$,

so that alternatively a balancing score for the problem is $b(X) = \{P^1(X), P^2(X)\}$.

For the other set of parameters, a finer score than the propensity score is always the 3-dimensional $b(X) = \{P^1(X), P^2(X), P^3(X)\}$:

When matching on the 1-dimensional propensity score, imposing common support can be done in terms of this scalar; when matching on the other balancing scores -> imposed on each element.

2. $E(Y^2 - Y^0 \mid D = j)$ for $j = 0, 2$

   The propensity score is $\frac{r(2,X)}{r(2,X)+r(0,X)}$,

   so that a balancing score for the problem is $b(X) = \{P^1(X), P^2(X), P^3(X)\}$.

3. $E(Y^2 - Y^1 \mid D = j)$ for $j = 1, 2$

   The propensity score is $\frac{r(2,X)}{r(2,X)+r(1,X)}$,

   so that a balancing score for the problem is $b(X) = \{P^1(X), P^2(X), P^3(X)\}$.

4. $E(Y^3 - Y^0 \mid D = j)$ for $j = 0, 3$

   The propensity score is $\frac{r(3,X)}{r(3,X)+r(0,X)}$,

   so that a balancing score for the problem is $b(X) = \{P^1(X), P^2(X), P^3(X)\}$.

5. $E(Y^3 - Y^1 \mid D = j)$ for $j = 1, 3$

   The propensity score is $\frac{r(3,X)}{r(3,X)+r(1,X)}$,

   so that a balancing score for the problem is $b(X) = \{P^1(X), P^2(X), P^3(X)\}$.

6. $E(Y^3 - Y^2 \mid D = j)$ for $j = 2, 3$

   the propensity score is $\frac{r(3,X)}{r(3,X)+r(2,X)}$,

   so that a balancing score for the problem is $b(X) = \{P^1(X), P^2(X), P^3(X)\}$.

Table 4: Appendix B: Summary Statistics: NCDS Men

Variable	3639 Observations	
	Mean	(S.D.)
Real log hourly wage 1991	2.040	(0.433)
Qualifications:		
O levels or equivalent	0.821	(0.383)
A levels or equivalent	0.548	(0.498)
Higher Education	0.283	(0.451)
White	0.969	(0.173)
Maths ability at 7:		
5th quintile (highest)	0.212	(0.408)
4th quintile	0.190	(0.392)
3rd quintile	0.185	(0.389)
2nd quintile	0.158	(0.365)
1st quintile (lowest)	0.141	(0.348)
Reading ability at 7:		
5th quintile (highest)	0.165	(0.371)
4th quintile	0.187	(0.390)
3rd quintile	0.188	(0.391)
2nd quintile	0.179	(0.383)
1st quintile (lowest)	0.166	(0.372)
Ability at 7 missing	0.115	(0.319)
Comprehensive school 1974	0.468	(0.499)
Secondary modern school 1974	0.162	(0.368)
Grammar school 1974	0.099	(0.299)
Private school 1974	0.052	(0.222)
Other school 1974	0.018	(0.134)
Father's years of education	7.270	(4.827)
Father's education missing	0.172	(0.377)
Mother's years of education	7.342	(4.606)
Mother's education missing	0.159	(0.366)
Father's age 1974	43.171	(13.736)
Father's age missing	0.075	(0.263)
Mother's age 1974	41.475	(10.864)
Mother's age missing	0.049	(0.216)
Father's social class 1974:		
Professional	0.044	(0.205)
Intermediate	0.145	(0.352)
Skilled non-manual	0.076	(0.265)
Skilled manual	0.297	(0.457)
Semi-skilled non-manual	0.010	(0.098)
Semi-skilled manual	0.095	(0.293)
missing	0.106	(0.308)
Mother employed 1974	0.513	(0.500)
Number of siblings	1.692	(1.789)
Number of siblings missing	0.106	(0.308)
Number of older siblings	0.821	(1.275)
Father's interest in edn:		
Expects too much	0.013	(0.114)
Very interested	0.252	(0.434)
Some interest	0.215	(0.411)
Mother's interest in edn:		
Expects too much	0.032	(0.175)
Very interested	0.344	(0.475)
Some interest	0.354	(0.478)
Bad finances 1969 or 1974	0.159	(0.365)
Region 1974:		
North Western	0.100	(0.300)
North	0.070	(0.256)
East & West Riding	0.079	(0.270)
North Midlands	0.072	(0.258)
Eastern	0.073	(0.261)
London & South East	0.143	(0.350)
Southern	0.057	(0.232)
South Western	0.061	(0.240)
Midlands	0.088	(0.283)
Wales	0.054	(0.227)
Scotland	0.096	(0.295)