NBER WORKING PAPER SERIES

ECONOMETRIC METHODS FOR RESEARCH IN EDUCATION

Costas Meghir
Steven G. Rivkin

Working Paper 16003
http://www.nber.org/papers/w16003

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
May 2010

This paper has been prepared for the Handbook of Education. We thank Rick Hanushek, Jim Heckman and Jeremy Lise for comments and discussions. We also thank Zohar Perla for her excellent research assistance. Costas Meghir thanks the ESRC under the Professorial Fellowship RES-051-27-0204 and the ESRC Centre for Microeconomic Analysis of Public Policy at the Institute for Fiscal Studies for funding this research. Rivkin would like to thank the Smith Richardson Foundation, the Spencer Foundation, the Hewlett Foundation, and the Packard Humanity Institute for supporting his work on the modelling of student achievement and teacher value-added. Responsibility for any errors is ours alone. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Econometric Methods for Research in Education
Costas Meghir and Steven G. Rivkin
NBER Working Paper No. 16003
May 2010
JEL No. C1,C14,H31,H52,I21,J24,J31

ABSTRACT

This paper reviews some of the econometric methods that have been used in the economics of education. The focus is on understanding how the assumptions made to justify and implement such methods relate to the underlying economic model and the interpretation of the results. We start by considering the estimation of the returns to education both within the context of a dynamic discrete choice model inspired by Willis and Rosen (1979) and in the context of the Mincer model. We discuss the relationship between the econometric assumptions and economic behaviour. We then discuss methods that have been used in the context of assessing the impact of education quality, the teacher contribution to pupils' achievement and the effect of school quality on housing prices. In the process we also provide a summary of some of the main results in this literature.

Costas Meghir
Department of Economics
University College London
Gower Street
London WC1E 6BT ENGLAND
c.meghir@ucl.ac.uk

Steven G. Rivkin
Amerst College
Department of Economics
P.O. Box 5000
Amherst, MA 01002-5000
and NBER
sgrivkin@amherst.edu

Contents

1 Introduction

The rising return to schooling and growing evidence in support of education as a primary determinant of economic growth has elevated the importance and visibility of research on human capital formation including both the determinants of enrolment and attainment and the determinants of education quality. Such research must address complications introduced by the myriad and inter-related decision making processes of families, teachers, administrators and policy makers. A variety of methods have been used to identify causal relationships, ranging from structural models based on utility maximisation to experimental and quasi-experimental approaches, yielding a growing body of often contradictory evidence.

Although the various approaches differ in the degree to which theoretical models of decision-making underlay the empirical specifications, the simple dichotomy between structural approaches on the one hand and experimental or quasi-experimental on the other does not hold up in most applications. As we highlight throughout the chapter, the interpretation, the usefulness and even the identification of estimates typically relies implicitly if not explicitly on a set of assumptions about underlying behaviour. Importantly, the introduction of heterogeneity in treatment effects magnifies the importance of such assumptions.

Rather than dividing this chapter by methods, we divide it into two parts largely in parallel to the division of research on human capital formation into quantity (years of schooling) and quality. The first part focuses on the estimation of wage equations and the return to schooling, framed by a Roy Model of education decision-making, originally suggested by Willis and Rosen (1979), that incorporates heterogeneity in returns to schooling across both individuals and three levels of schooling and allows for comparative advantage in the sense that individuals need not be best at both education and the labour market. This model not only provides a flexible description of the process through which human capital is acquired through schooling, but it also provides a

framework for discussing the identification issues that arise when estimating education effects on wages with empirical methods not based directly on models that articulate the full structure of the process of human capital acquisition.

We then turn to the various methods used to estimate education effects on wages, highlighting the restrictions that must be fulfilled in order to generate consistent estimates of the return to schooling. The natural starting point is the estimation of the full structural model. We describe its estimation based both on maximum likelihood and on newly developed simulation methods.

Next we describe methods that can be used to estimate Mincer wage equations in which education is taken to be a continuous variable. First, we discuss approaches to identification under the assumption that the return to schooling is constant both with respect to the years of education and across individuals. We then permit the wage return to schooling to vary with the level of education and discuss the use of non-parametric IV estimators when the shape of the relationship between wages and years of education is not known. Next we allow for heterogeneity across individuals and consider general non-separable models with respect to years of education and unobserved heterogeneity. This takes us to the frontier of research as far as this class of models is concerned. Importantly, it becomes clear that identification and interpretation of results depends on the nature of schooling choice. Indeed the implied restrictions on economic behaviour, required for identification are quite stringent and can be interpreted as restrictions on the information possessed by the individual when making education choices.

Following discussion of the Mincer wage equation we return to the Roy model with potentially unordered schooling choices and discuss approaches to identification of the effects of education on wages. There we discuss conditions that allow identification "at infinity", an argument which depends on the availability of enough continuous instruments, such that there are sets of values of these instruments where the individuals facing such values make a choice of a particular education level with probability one.

In such sets there is no selection. If the instruments are independent of the unobservables in the wage equation and under some further conditions identification is achieved. However, it is unlikely that such an identification strategy has much empirical significance. At this point we either need to acknowledge the need for further parametric assumptions, beyond those implied from theory and beyond the standard IV assumptions or we need to resort to set identification. In this case we focus on bounding the distribution of wages for each education group, rather than attempting to obtain point estimates. The various assumptions used in the point identification approach, such as instrument exclusion can be used here; some can even be relaxed to allow for instruments to shift the distribution of wages in one direction (monotonicity).[1] The approach can be very fruitful, because it provides an intermediate position where a number of theoretical restrictions are used, making the estimates interpretable within a broad theoretical framework, while not using auxiliary assumptions that are both controversial and do not obtain from theory.

The final section of Part I discusses the problem introduced by missing wages for labour force nonparticipants. We compare a fully structural approach that in practice will have to rely on assumptions beyond those implied by theory and an approach based on bounds.[2] Allowing for both endogenous education choice and endogenous labour force participation does make identification more stringent. At the same time bounds are also less informative. Of course ignoring these issues poses serious problems with interpretation. Identification will have to rely on exploiting the restrictions from theory as well as further assumptions, such as those specifying the distribution of unobservables. Despite the shortcomings of having to make assumptions that do not relate directly to theory, the approach that ignores these issues leads to results that have limited interpretation. Moreover, while most identification theorems proposed hitherto rely on identification at infinity arguments, it is possible that further progress can be

[1] See Manski (1994)

[2] See Blundell, Gosling, Ichimura and Meghir (2007)

made by exploiting further restrictions from theory; one possibility is to explore the use of restrictions from other related decisions. The potential for this can be seen when comparing the identification of pure discrete choice models with those who combine discrete choice with continuous outcome variables, such as the education and wages model we discuss.[3] Moreover when considering bounds it is clear that identification can obtain without an "identification at infinity" argument. Characterising the underlying behavioural conditions for this would be an important advance.

We start the section below by justifying the idea of a Roy model of education and wages by suggesting that each education level may correspond to a different input in production, these inputs not being perfectly substitutable for each other. We argued that in an economy with changing supply of educated workers of differing levels, the relative wages and the returns to education will change over time. We also refer to evidence that demonstrates the importance of such considerations. In view of this, we close this section by discussing the implications or policy of placing education choice and wages within a general equilibrium framework. A number of authors have shown that without such a framework it is very difficult to design and think of policy.[4] This again emphasises the need for a model not only for the interpretation of the estimates but also for understanding what the estimates imply for policy.

Part 2 turns to empirical methods used in research on school and teacher quality. Although Part 2 begins with the presentation of an education production function model and discussion of the multiple levels of choices that determine the matching of students, teachers, and schools, this model does not provide the unifying framework of the Roy Model in Part 1. Given the range of issues covered in research on education quality, we believe it to be more productive to focus on conceptual frameworks tailored to specific issues. We do, however, begin Part 2 with general discussions of the housing choice and the dynamics of the process of knowledge acquisition.

[3] Contrast Magnac and Thesmar (2002) with Heckman and Navarro (2007) for example.

[4] See Heckman, Lochner and Taber (1998), Lee (2005), Lee and Wolpin (2006) and Gallipoli, Meghir and Violante (2008).

We divide Part 2 into four sections corresponding to different research areas that have received substantial attention in recent years, and within each we juxtapose various empirical to estimation. Specifically, we focus on a small number of papers on class size, teacher quality, competition and accountability, and capitalisation of school quality into house prices. The proliferation of administrative and survey data in recent years has facilitated research on these and other education topics, and we have selected papers that vary by both type of data and empirical method. The methods include controls for observables, instrumental variables, regression discontinuity as a special case of IV, use of random lotteries as a special case of IV, difference-in-differences, fixed effects with large administrative data sets, and the use of data generated by experiments.

As in the first section we highlight the inter-relationship among the structure of underlying choices, treatment effects, identification conditions, and meaning of the estimates. Estimators differ according to the assumptions required for identification and assumptions concerning the distribution of treatment effects along various dimensions, though most of these estimators do not come from behavioural models that predict the structure of treatment effects. Nonetheless, we focus on the inter-dependencies among underlying behaviour, identification conditions and interpretation throughout this discussion.

In the case of research on class size, we first present a model of education production based on Lazear (1999) that highlights potential dimensions over which the benefits of smaller classes might vary and then evaluate a series of different estimators with that framework as a backdrop. In addition to describing the methods and identifying assumptions, this compares estimators according to the degree to which they capture class size related general equilibrium effects on the quality of instruction; some estimators capture cross-sectional differences in teacher quality associated with smaller classes, some capture changes over time in state average teacher quality, while others isolate the ceteris paribus effect of smaller classes.

The description of research on the semi-parametric estimation of the variance in teacher quality begins with a discussion of the behavioural responses of families, teachers, and administrators that complicate estimation. It then describes different approaches to estimation of the variance and to accounting for sorting both between and within schools.

The discussion of research on housing market capitalisation focuses narrowly on boundary fixed effects estimators (see Chapter # for a comprehensive treatment of this issue). Two of the three papers adopt reduced form approaches, while the third develops a discreet choice framework with which to model housing choice and heterogeneity in the preferences for both school quality and peer characteristics.

The papers on choice and accountability examine different types of incentives including competition from public and private schools and state accountability systems. Non-random family take-up of choice options complicates estimation of choice effects, while the non-random distribution of accountability system adoption date and character complicate efforts to identify accountability system impacts. Each paper takes a different approach to account for unobserved differences across families and states.

The chapter concludes with a synthesis of the key commonalities and differences between work on the return to schooling and on school quality and the ways in which the proliferation of administrative and survey data affect the structure of empirical analyses. It highlights remaining challenges and areas for additional work.

2 Wage equations and the returns to education

The aim of this section is to discuss the estimation of the returns to education. We place wages and education choice within a simple competitive general equilibrium framework. In this model forward looking individuals choose an education level; each level is associated with its own wage process. The allows us to clarify the notion of returns to education and to discuss clearly identification issues. Once we have defined the

model of education choice and wages we digress to the simpler world of Mincer wage equations and discuss identification issues within that context. In a simple version of that model, as shown by Mincer (1958, 1974), the returns to education can be estimated directly from the relationship of wages and education, because the only cost to education is the opportunity cost - there are no direct costs such as fees. Moreover in many contexts we may be interested in the impact of education on wages in and of itself. Following the discussion of identification of Mincer type models we return to the identification issues of the dynamic model we originally introduced. One of the main themes of this section will be the extent to which we can interpret estimates on the wage returns to education without saying much about the process of education choice.

2.1 Pricing of human capital

Our starting point is that production involves k types of human capital; the type of human capital that an individual possesses is determined by the level of education they have attained. Here the levels considered will be statutory schooling, high school and college. The wage received by an individual will we the product of the aggregate price of her type of human capital (W_{kt}) with the amount she brings to the market (h_{kit}), say $w_{kit} = W_{kt}h_{kit}$, where k denotes the type of human capital and h_{kit} is the amount in efficiency units that individual i possesses in time period t.[5] The literature on modelling wages concentrates on understanding how to model the constituents of h_{kit}, which will be a function of education, ability and possibly experience, age and other factors that enhance individual productivity and skills. Thus the returns to education in a competitive economy will depend both on how each education level is priced in the market (i.e. how W_{kt} is determined) and how education contributes to the formation of h_{kit}. The way pricing may change in the future is a source of aggregate uncertainty,

[5] See for example Heckman, Layne-Farrar and Todd (1996).

while possible future shocks to human capital h_{kit} will be a source of idiosyncratic uncertainty.

Consider first pricing in a competitive market. Suppose there are three levels of education: less than high school (S^L), high school (S^H) and college (S^C). Now suppose we can represent aggregate production by the function

$$Q_t = A_t \left[\delta_S H_{St}^{\rho} + \delta_H H_{Ht}^{\rho} + \delta_C H_{Ct}^{\rho}\right]^{\alpha/\rho} K_t^{1-\alpha}$$

with $\delta_i > 0$, $\delta_C + \delta_H + \delta_S = 1$, $\rho \leq 1$, $0 \leq \alpha \leq 1$ and H_{jt} is the sum of total human capital employed of type j in period t. In efficiency units this is $H_{jt} = \sum_{i=1}^{N} h_{kit}$, where h_{kit} is the human capital of type k supplied by individual i in period t. In a competitive equilibrium prices satisfy

$$\frac{W_{jt}}{W_{St}} = \frac{\delta_j}{\delta_S}\left(\frac{H_{jt}}{H_{St}}\right)^{\rho-1} \tag{1}$$

Thus the relative prices of the two types of human capital will vary depending on the ratio of demands $\frac{H_j}{H_S}$, so long as $\rho < 1$. The source of aggregate uncertainty are changes in the relative value of the δs, such as skill biased technical change. When $\rho = 1$ the inputs are perfectly substitutable and the relative prices are invariant to changes in $\frac{H_j}{H_S}$. The relative pay can change however in response to changes in technology as expressed by changes in $\frac{\delta_j}{\delta_k}$. The resulting wage equation for individual i who has obtained educational level k can then be represented as

$$\ln w_{kit} = \ln W_{kt} + \ln h_{kit} \tag{2}$$

When $\rho = 1$, and with no changes in technology this simplifies to

$$\ln w_{it} = \ln W_t + \ln h_{it} \tag{3}$$

which is now common across education groups. Wage equations based on years of

education only, without regard to the type of education received can be interpreted through, and is inspired by the work of Mincer.

One way of estimating the parameters is to use 1 and write

$$\log \frac{W_{jt}}{W_{St}} = \log\left(\frac{\delta_j}{\delta_S}\right) + (\rho - 1)\log\left(\frac{H_{jt}}{H_{St}}\right) + \varepsilon_{jS} \tag{4}$$

To implement this we need measures of $\frac{W_{jt}}{W_{st}}$ and of $\frac{H_{jt}}{H_{st}}$. These can be obtained as follows. First we estimate wage equations for each level of schooling. The time dummies in the equations for level S is a measure (up to an additive constant) of $\log W_{St}$. an estimate of individual human capital can then be obtained by using the relationship

$$h_{kit} = \exp(\log w_{kit} - \log W_{kt})$$

Aggregating the estimate of this quantity and assuming the sample is representative (or grossing it up with the sample weights) allows us to construct $\log\left(\frac{H_{jt}}{H_{st}}\right)$. If the errors in 4 relate to changes in technology then human capital is endogenous. One strategy is to instrument $\log\left(\frac{H_{jt}}{H_{st}}\right)$ by the total number of people of these two skill levels, whether they are working or not. This uses the medium term availability of the resource as the instrument. An alternative would be to use dynamics, i.e. lags in human capital as instruments.

The empirical evidence suggests that $\rho < 1$ and hence that different levels of human capital are not perfectly substitutable. Katz and Murphy (1992) and Heckman, Lochner and Taber (1998) both estimate the elasticity of substitution ($\frac{1}{1-\rho}$) between unskilled and skilled workers to be about 1.4, although alternative estimates suggest even lower values. Gallipoli, Meghir and Violante (2008) estimate the elasticity of substitution between statutory schooling, high school and college for the US using instrumental variables. They find that the same elasticity of substitution can be imposed across all

pairwise comparisons; they estimate this to be between 1.5 and 2.[6]. It is safe to say that the consensus in the literature is that different types of human capital are not perfectly substitutable for each other.

2.2 A model of education choice and wages

Given the discussion above it follows that we should model education as choosing a specific level; we should then estimate lifecycle wage profiles within each sector. This is similar to the Roy (1951) model and the empirical basis for this has been formulated by Willis and Rosen (1979). Keane and Wolpin (1997) have taken this further, allowing also for occupational choice and modelling the entire career.[7]

This model provides a framework for estimating the returns to education and discussing the identification issues that arise when we attempt to estimate the wage returns with simpler methods and without articulating the entire structure of the model. Indeed we will argue that the identification of wage returns to education are intricately linked with the underlying model of education choice and that in most cases these cannot be separated.

Consider an individual who has just completed statutory schooling and will decide on completing high school and then on whether to complete college. We will take these as two sequential decisions, to which the individual needs to commit successively. Once schooling is over the individual works and earns depending on the level of

[6]They consider less than high school, high school, some college and college. They could not reject the restriction of one elasticity of substitution.

[7]Variations of this model have been used by Attanasio, Meghir and Santiago (2009) where only the discrete education choice is modelled and Gallipoli, Meghir and Violante (2008) who model educational choice, wages, labour supply and intergenerational transfers. Heckman and Navarro (2007) use a version of this model to analyse identifiability of returns to education and of the dynamic discrete choice models.

schooling received. The earnings function for each level of education is given by

$$\begin{aligned}
\log w_i^S &= m^S(age, X_i) + \tau_i^S + \varepsilon_{it}^S \qquad \text{Statutory Schooling} \\
\log w_i^H &= m^H(age - e^H, X_i) + \tau_i^H + \varepsilon_{it}^H \qquad \text{High School} \\
\log w_i^C &= m^C(age - e^C, X_i) + \tau_i^C + \varepsilon_{it}^C \qquad \text{College}
\end{aligned} \tag{5}$$

where e^H and e^C represent the additional years of schooling, over and above the statutory ones needed to obtain a high school and college degree respectively and where X_i are observable characteristics that influence individual earnings. It would be straightforward to include age and other time varying characteristics so long as we made the assumption of perfect foresight. At a later section we show how the model can be generalised to include endogenous experience and nonparticipation. We have abstracted from the aggregate fluctuations in the prices of human capital, but in this simple environment allowing for aggregate shocks is relatively straightforward.

These earnings functions include a number of important features: first earnings growth with respect to potential experience, i.e. age minus years of education differ depending on the level of education, allowing for the possibility of complementarity between education level and age. Second each individual has a sector specific level of ability (τ_i^0, τ_i^H, τ_i^C). This implies heterogeneous returns to education, with respect to unobservables (as well as possibly through observables). We will assume here that this heterogeneity is in the individual's information set when making the education choice. This assumption is not innocuous and does affect the way we estimate the model. Third, the stochastic structure of earnings will differ across groups. For example the variance or persistence of shocks may differ depending on the chosen sector. This is particularly important when considering models that allow for risk aversion. The model has abstracted from other important issues, such as endogenous experience

and non-participation.

Now consider the flow of utility. We assume that individuals incur observable monetary costs $\kappa^H(Z^H)$ and $\kappa^C(Z^C)$ for high school and college, respectively. These will include fees, cost of books, transport costs etc. and may depend on observable characteristics $Z_i = \{Z_i^H, Z_i^C\}$. Individuals will also incur unobservable costs per unit of time spent in education, which we can interpreted as effort. These we will assume are heterogeneous and we denote them by F_i. Taking all this into account the flow utility for high school and for college is respectively

$$\begin{aligned} u_i^H &= \kappa^H(Z_i^H) + F_i + \nu_i^H \quad \text{High School} \\ u_i^C &= \kappa^C(Z_i^C) + \alpha F_i + \nu_i^C \qquad \text{College} \end{aligned} \tag{6}$$

where the coefficient α reflects the fact that effort and time spent in college can be different than in high school and where the ν^H and ν^C represent random shocks to the costs of education. The high school utility ν_i^H shock is revealed when the individual needs to decide to go to high school or start working. Similarly the college shock is revealed after high school when the college decision needs to be made. The assumed timing of such shocks is critical for the model: the fact they are revealed sequentially, implies that the decision to continue to the next level will include an option value to continue further; thus as we shall see, attending high school has value for the earnings that are expected and because of the option of attending college, whose value will depend on the shock that has not been realised yet. We will simplify the model further by assuming that once the individual has dropped out of the formal education system they no longer can return and they work until a fixed retirement age of say 60. In a more detailed model we would allow for endogenous retirement or at least the recognition that retirement age differs by education group and is probably higher for the more educated. The expected utility from working having achieved education level S, H or

C respectively, as viewed at the time the decision will be made is

$$V_i^{WS} = w_{i0}^S + E\sum_{t=e^S+1}^{60}\beta^t w_{it}^S = \exp(\tau_i^S)A^S(X_i) \qquad \text{Statutory School Earnings}$$

$$V_i^{WH} = w_{ie^H}^H + E\sum_{t=e^H+1}^{60}\beta^t w_{it}^H = \exp(\tau_i^H)A^H(X_i) \qquad \text{High School Earnings} \qquad (7)$$

$$V_i^{WC} = w_{ie^C}^C + E\sum_{t=e^C+1}^{60}\beta^t w_{it}^C = \exp(\tau_i^C)A^C(X_i) \qquad \text{College Earnings}$$

where β is the personal discount and $A^J(X) = \sum_{t=e^J+1}^{40}\beta^t \exp\left[m^H(t-e^J, X_i)\right]$, $J = S, H, C$, with e^i being the age at which education level i is completed. The expectation is taken over future wage shocks. Now consider the decision process for someone who just completed statutory schooling. The value of attending high school will be given by the sum of the current costs of schooling (u_i^H) and the option value of either going to college or starting work as a high school graduate. Thus the value of attending high school is given by

$$V_i^H = \kappa^H(Z_i^H), + F_i + \nu_{it}^H + \beta E \max\left\{V_i^C, V_i^{WH}\right\}$$

$$= \tilde{V}_i^H + \nu_i^H$$

where the expectation is over all future wage shocks and the shock to the cost of attending college v_i^C. The value of attending college is given by

$$V_i^C = \kappa^C(Z_i^C) + \alpha F_i + \nu_i^C + \beta E V_i^{WC}$$

$$= \tilde{V}_i^C + \nu_i^C$$

The first decision is to attend high school or not. The decision rule is

$$\text{Attend High School} \Longleftrightarrow V_i^H > V_i^S \tag{8}$$

If the individual does not attend high school then they enter the labour market until retirement. If they do attend, in the next period they need to decide whether to continue with college. The decision rule is again

$$\text{Attend College} \Longleftrightarrow V_i^C > V_i^{WH}. \tag{9}$$

At the point of making the education decision the individual is assumed to know the experience profiles of each education sector as well as the permanent heterogeneity components τ_i^C, τ_i^H and τ_i^S. They also know the costs of education (both direct and effort costs). Interestingly, this model allows for comparative advantage (as in Willis and Rosen, 1979)) in the sense that individuals need not be best at both education and the labour market: they may be very good at the medium skill labour market (high τ_i^H) and not so good at education (low F_i) or perhaps a negative αF_i. Thus both the mechanism of selection into different education levels and the resulting relationship between education and unobserved components of wages are complex and not necessarily in an easy to predict direction.

To estimate the model or use it for simulations we must first solve it, i.e. compute expressions for the value functions that will allow us to implement the decision rules 8 and 9. In the context of this particularly simple model this is easy to do, but in more complex models it can be computationally time consuming.

Given a set of parameters and given the distribution of unobserved heterogeneity the discounted value of further schooling and/or work can be computed by projecting

earnings forward based on the wage equation for the relevant education sector. Thus for example

$$V_i^{WS} = w_{i0}^S + E \sum_{t=e^S+1}^{60} \beta^t w_{it}^S$$

The next step involves a conditional expectation with respect to the distribution of the shocks. Here we either need to make an explicit distributional assumption about the shocks $\nu_i^C - \nu_i^H$ or we need to see if such a distribution can be identified nonparametrically. In a parametric context these could be assumed normal for example; since the decision is discrete we normalise the variance to be one. The distribution is first needed to write down the expected value function for the next period, which involves the future optimal decision to continue schooling:

$$E \max \left\{ V_i^C, V_i^{WH} \right\} = E(V_i^C | \nu_i^C - \nu_i^H > \tilde{V}_i^{WH} - \tilde{V}_i^C) \Pr \left(\nu_i^C - \nu_i^H > \tilde{V}_i^{WH} - \tilde{V}_i^C \right) +$$

$$+ E(V_i^{WH} | \nu_i^C - \nu_i^H < \tilde{V}_i^{WH} - \tilde{V}_i^C) \Pr \left(\nu_i^C - \nu_i^H < \tilde{V}_i^{WH} - \tilde{V}_i^C \right)$$

The missing components to evaluate this expression are $\tilde{V}_i^{WH} = E \sum_{t=e^H+1}^{60} \beta^t w_{it}^H$ and $\tilde{V}_i^C$. Given a wage equation and a discount factor we can easily compute $\tilde{V}_i^{WH}$, which is the present discounted value of mean earnings.

In a model that is linear in earnings the value beyond working life can be set to zero, without loss of generality. However, a practical difficulty is that we may be missing wages for older individuals, during working life and this needs to be accounted for. Suppose we observe individuals up to some age (it could differ across individuals). Then we either need to assume how earnings will evolve beyond that age or we need to introduce a terminal value function whose parameters will be estimated alongside the remaining parameters of the model. If we have enough data to estimate the age-earnings profile only up to a specific age T the terminal value function will be a function of the state variables at that age as well as of observed X_i and unobserved characteristics

unobserved heterogeneity. In our simple model the state is just the education level . Thus we can specify

$$V_i^{JT} = \exp(\gamma_0^J(X_i) + \gamma_1^J(X_i)\tau_i^J) \text{ for } J = S,H,C \tag{10}$$

where we have used the exponential to restrict the function to be positive. Then we can compute $\tilde{V}_i^{WJ}$ as follows

$$\tilde{V}_i^{WJ} = \sum_{t=s+1}^{T_i} \beta^t \hat{w}_{it}^J + \beta^{T_i+1} V_i^{JT}$$

where $\hat{w}_{it}^J$ is the expected wage conditional on unobserved heterogeneity τ^J . Thus now for a given set of parameters, we can compute all future values required to construct the probability of attaining a particular level of education, conditional on τ.

The average lifecycle returns to college *vis a vis* high school from the perspective of someone who has just completed high school is

$$\Delta^{C/H} = E\left[\frac{V^C - V^{WH}}{V^{WH}}\right]$$

where the expectation is taken with respect to the unobservables. This measure includes both the wage gains of going to college over the lifecycle as well as accounting for the individual direct and opportunity costs of education. This measure will allow for the effects of education on other dimensions of behaviour, such as expected endogenous and exogenous spells out of work, which can be elements of more complex models.[8]

In this model the returns to education are heterogeneous, depending on the unobserved components of wages and educational costs. The individual wage returns to

[8] see for example Adda, Dustmann, Meghir and Robin (2009)

college, at any specific point in the lifecycle take the form

$$\Delta W_i^{C/H} = m^C(age - e^C) - m^H(age - e^H) + \tau_i^C - \tau_i^H + \varepsilon_{it}^C - \varepsilon_{it}^H$$

Since the wage returns[9] to education are heterogeneous in this model there are many different concepts of such returns.[10] The average wage return to one level of education L relative to another L' (Average Treatment Effect) at a specific age is given by

$$\Delta_{ATE}^{L/L'}(age, X_i, L, L') \equiv$$

$$E(\ln w_i^L - \ln w_i^{L'} | X_i) = m^L(age - e^L, X_i) - m^{L'}(age - e^{L'}, X_i)$$

The return relative to L' for those who chose $J = L$ (Average Treatment on the Treated) is given by

$$\Delta_{ATT}^{L/L'}(age, Z_i, X_i, L, L') \equiv E(\ln w_i^L - \ln w_i^{L'} | X_i, J = L)$$

$$= m^L(age - e^L, X_i) - m^{L'}(age - e^{L'}, X_i) + \tag{11}$$

$$\Big[E(\tau_i^L | age, Z_i, X_i, J_i = L) - E(\tau_i^{L'} | age, Z_i, X_i, J_i = L)\Big]$$

Note that in 11 the last expression is the average labour market ability relating to high school education for those *who chose to attend college*. In all cases, estimation of the returns of interest (e.g. ATE or ATT or the entire distribution) will require estimating some aspect at least of the distribution of the τ^J. Just to emphasise this point, if we ignore this issue and we just compare wages across sectors the estimate of the returns

[9] Henceforth we will refer loosely to the "returns to education" as the effect of education on wages.
[10] See Heckman, LaLonde and Smith (1999)

we will obtain will have the form

$$D^{L/L'} = \left[m^L(age - e^L, X_i) - m^{L'}(age - e^{L'}, X_i) \right] +$$

$$+ \left\{ E(\tau_i^L | age, Z_i, X_i, J_i = L) - E(\tau_i^{L'} | age, Z_i, X_i, J_i = L') \right\}.$$

While the first term in square brackets is indeed $\Delta_{ATE}^{L/L'}(age)$ and reflects the gains obtained as a result of attending college over high school, the term in $\{\}$ brackets represents the differences in composition between the two groups. This differs from the expression is the square brackets in 11, which represents the average ability for education level L of those who chose L minus the average ability for level L' *of the same group of individuals*, i.e. those who chose L.

2.3 Estimation

The model described above has left a number of objects unspecified. These include the functional forms for the direct costs of schooling $\kappa^J(Z_i^J)$, $J = H, C$, the functional form for the wage equations and the distribution of preferences and wages induced by the random vector $(F_i, \tau_i^0,\ \tau_i^H,\ \tau_i^C, v^H, v^C, \varepsilon_i^0,\ \varepsilon_i^H,\ )$. The identifiability of such a model is an important point of discussion. Indeed Magnac and Thesmar (2002) have shown that the discrete choice model, without any direct link to outcomes such as wages is underidentified.[11] In their context they prove that we can identify the static utilities (here direct costs of education) if we fix the discount rate, the distribution of preferences and the utility of a reference choice. The implication is that in a discrete choice model forward looking dynamics have little empirical content; for example we cannot distinguish nonparametrically between a forward looking model and a static one (discount rate zero) without further restrictions. However, this framework is perhaps

[11] See also Rust (1994)

asking too much from the data and it is perhaps not too surprising that with just discrete decisions and no other restrictions we cannot identify much. Heckman and Navarro (2007) argue that using the cross equation restrictions between educational choice and wages, as implied by a model where educational choices depend on labour market gains, and putting some (factor) structure on the distribution of unobservables is crucial for identification, although not sufficient. We discuss these identification issues below. Here we address estimation in a fully parametric context. That is we specify all missing functional forms, including the distribution of unobservables, up to an unknown finite set parameters.

There are numerous ways to implement estimation; for particularly complex models a number of simulation approaches have been developed, including simulated method of moments and indirect inference.[12] Here we describe maximum likelihood because the model we presented is relatively simple. We briefly discuss the implementation of simulation estimators for this type of model below.

The estimation approach suggested by Rust (1987), known as Nested Fixed Point (NFP) algorithm involves starting at some initial parameter vector; solving the model to obtain the future value functions V^C, V_i^{WH} at all possible values of observables and unobservables; followed by the evaluation of the the likelihood function. Once this is evaluated at the parameter vector an update of the parameters can be found based on a suitable optimisation algorithm, such as Gauss-Newton; when the updated parameter vector has been obtained the process starts again with the solution of the model and so on until convergence. The model will also have to be solved in intermediate steps so as to be able to compute the derivatives of the likelihood during the Gauss Newton iterative process or other derivative based method. It is thus crucial to solve the model in a computationally efficient way.

Su and Judd (2008) note than in many cases the NFP algorithm can become ex-

[12] See McFadden (1989), Pakes and Pollard (1989) and Gourieroux, Monfort and Renault (1993) amongst others.

tremely time consuming and possibly infeasible because of the huge number of times it needs to solve the full dynamic programming problem. They propose an alternative approach based on mathematical programming subject to equilibrium constraints that simultaneously solves for the value function and estimates the parameters θ. Effectively, they treat the unknown value functions as parameters to be estimated and define the link between the value functions and the structural parameters as a set of nonlinear constraints on the parameters. Once set up in this way any standard optimisation algorithm can be used. They find one obtains both speed and accuracy gain in the Rust type problem.

To see how the model can be solved we start by constructing the probabilities of educational attainment *conditional* on unobserved heterogeneity.

a. An individual is observed having just statutory education. They are then observed for T_i periods in the labour market. The probability of this level of schooling is

$$P_i^S \equiv \Pr(\text{Statutory Schooling}) = \Pr\left(v_i^H < EV_i^{WS} - \kappa^{H\prime} Z_i^H - F_i - \beta E \max\left\{V_i^C, V_i^{WH}\right\}\right) \tag{12}$$

This probability depends on the costs of high school, on the stream of earnings resulting from work at this level of qualification, as well as on the benefits of continuing education into college, which is expressed as a comparison between the benefits of college and the stream of future incomes from high school. Interestingly, even if the benefits from high school may be low for a particular individual, because say τ_i^H happens to be very low they may still choose to attend high school because their benefits from college may be very high. This illustrates why final educational attainment cannot in general be represented as an ordered discrete choice regression model, which will have implications below for the identification and estimation of the wage returns to education.[13] From an economic point of view this highlights the notion that each education level

[13] See Cameron and Heckman (1998)

may represent a different sector and individuals may have a comparative advantage for a more advanced sector, without being particularly good at the intermediate level.

Now consider what is the probability of observing someone completing high school and then going on to work. They prefer to continue after statutory schooling and stop after high school. To write down the probability of this event we will assume that the schooling shocks v^J, $J = S,H,C$ are independent over time and that all the dependence in the sequential decisions comes from F_i. Then the probability of high school completion with no further college is

$$P_i^H \equiv \Pr(\text{High School and not College}) =$$

$$\left[1 - \Pr\left(v_{it}^H < \beta E V_i^{WS} - \kappa^{H\prime} Z_i^H - F_i - \beta E \max\left\{V_i^C, V_i^{WH}\right\}\right)\right] \times$$

$$\Pr\left(v_{it+1}^C - v_{it+1}^H < \kappa_i^{H\prime} Z_i^H - \kappa_i^{C\prime} Z_i^C + (1-\alpha)F_i + \beta\left[E \max\left\{V^C, V_i^{WH}\right\} - E V_i^{WC}\right]\right) \tag{13}$$

Finally, the probability of completing college can be written as

$$P_i^C \equiv \Pr(\text{College}) =$$

$$\left[1 - \Pr\left(v_{it}^H < \beta E V_i^{WS} - \kappa^{H\prime} Z_i^H - F_i - \beta E \max\left\{V_i^C, V_i^{WH}\right\}\right)\right] \times$$

$$\left[1 - \Pr\left(v_{it+1}^C - v_{it+1}^H < \kappa_i^{H\prime} Z_i^H - \kappa_i^{C\prime} Z_i^C + (1-\alpha)F_i + \beta\left[E \max\left\{V^C, V_i^{WH}\right\} - E V_i^{WC}\right]\right)\right] \tag{14}$$

This sequence of probabilities also illustrates the way that the distribution of characteristics (observable and unobservable) evolve as individuals progress through schooling: all else being equal individuals with the lowest values of F are the first to stop school. The selection that happens next very much depends on the relative importance of effort costs F for high school and college. If F is more important for high school than college ($\alpha < 1$) then individuals with the highest values of F will actually drop out and not go

to college. If on the other hand effort is more important for college, the persons with highest value of F will complete college and then enter the labour market.

Now suppose we observe earnings for an individual over T_i periods. The key complexity here relates to the stochastic structure of earnings, i.e. the way that the error terms ε evolve over time. In this context we will deal with the simplest case where the ε are all independently and identically distributed over time. However, other more realistic assumptions in the literature include cases where the ε are a random walk or have an autoregressive structure.[14]

We suppose that observations on earnings start immediately after full time education is completed. Thus if $t = 1$ for the first observation of someone who started working following statutory schooling, $t = e_H + 1$ for someone with a high school degree and $t = e_C + 1$ for someone with college. With i.i.d. errors the density of the sequence of earnings for the ith individual can be written as

$$L_i^J \equiv L(w_{is}...w_{iT_i}|J_i, \tau_i, X_i, \theta) = \prod_{t=e_i}^{T_i} g^J(w_i|t - e_i,\ X_i,\ \tau_i^{J_i}) \tag{15}$$

where J_i is the level of education achieved, g^J is the density of wages for those with education J, and θ is the vector of parameters. Putting all the pieces together the likelihood function for an individual who has followed education stream J and is observed working until age T_i is

$$L_i(w_{is}...w_{iT_i}, J_i|Z_i, \tau_i, \theta) = P_i^J(J_i|Z_i, F_i, \tau_i, \theta) L_i^J(w_{is}...w_{iT_i}|J_i, \tau_i, \theta) \tag{16}$$

All probabilities and wages depend on unobserved heterogeneity components. The distribution of unobservables needs to be estimated together with the rest of the parameters. The model includes four unobservables, namely $\tau_i = (\tau_i^S,\ \tau_i^H,\ \tau_i^C)$, and F which

[14] Adda et al. (2009) allow for a within firm random walk.

need to be integrated out. Integrating out the unobserved heterogeneity implies that we average over all possible values, using as weights the probability of each possible value. The weights, as well as the possible values of unobserved heterogeneity constitute unknown parameters of the model. The individual contribution to the likelihood now becomes

$$E_{F,\tau}\left[L_i(w_{is}...w_{iT_i},J_i|Z_i,\tau_i,\theta)\right]=\int_{F,\tau}P_i^J(J_i|Z_i,F_i,\tau,\theta)L_i^J(w_{is}...w_{iT_i}|J_i,\tau,\theta)dG(F,\tau)$$

where $G(F,\tau)$ is the four dimensional distribution of unobserved heterogeneity. Note that we are allowing for correlation between the factors.

A key complication in practice is that such high dimensional integrals can take a long time to compute and while this level of generality is very attractive, in that it allows very general sorting patterns into different education groups, the computational difficulty could make the whole problem prohibitive, particularly in the context of richer and more complex models including other decisions such as labour supply.

We can keep some of the advantages of the original specification and simplify the problem substantially by assuming that there are just two factors, one that enters education as before (F) and one that enters wages (τ). The effect of the unobservable in each of the wages is controlled for by a coefficient α^J ($J=H$, C). The coefficient on the wages for statutory schooling is normalised to one since τ is not observed.

Whatever the specifics of the distribution of unobserved heterogeneity, the sample loglikelihood for N individuals is

$$LogL=\sum_{i=1}^{N}\log\left\{\int_{F,\tau}P_i^J(J_i|Z_i,X_i,F,\tau,\theta)L_i^J(w_{is}...w_{iT_i}|J_i,X_i,\tau,\theta)dG(F,\tau)\right\} \qquad (17)$$

The estimation problem relates to obtaining estimates for the unknown parameters θ and the distribution $G(F,\tau)$ by maximising $LogL$ in 17.

2.3.1 Estimation by simulation

Many structural models are often too complex to estimate based on Maximum Likelihood. Since the seminal work of Lerman and Manski (1981), McFadden (1989) and Pakes and Pollard (1989) simulation approaches and in particular simulated methods of moments have offered a useful alternative that allow us to approach much more complex models.

The first step in simulated method of moments is to decide on a set of moments that can identify the parameters on the model. In our education choice model the proportions attending each level of education, mean wages and variance of wages by education, all conditional on the exogenous variables would be suitable moments. These can be estimated directly from the data. Denote these by $\hat{q}$. Given a value for the parameter vector θ,which includes the distribution of unobservables we can simulate education choices and lifecycle profiles of wages from the model. The same moments that were estimated from the data can now be constructed from the simulated data. Denote these simulated moments by $q^s(\theta)$, where s denotes the number of simulations. An estimate of the covariance matrix of the estimated moments is $\hat{\Omega}$. Then the simulated method of moments minimises the function

$$Q(D|\theta) = (\hat{q} - q^s(\theta))' \hat{\Omega}^{-1} (\hat{q} - q^s(\theta)) \tag{18}$$

with respect to θ, where D is the data used to compute the moments Precision will improve with the number of simulations S used to compute $q^s(\theta)$ as well as with the degree of overidentification, i.e. the number of moments over and above those needed to exactly identify the model.

A note of caution is called for: while maximum likelihood uses all the information and restrictions implied by the model, given the available data, method of moments do not. In linear models it is easy to see what moments identify the model; however in

highly nonlinear models choosing the set of moments that can identify the model may be difficult in practice. Moreover it is complicated if not intractable to check formally that the chosen moments identify the model: this would involve checking that the 2nd derivative matrix of the criterion function is negative definite around the optimum. Thus the cost of moving away from maximum likelihood is the lack of clear rules for choosing the right moments to match.

Returning to the estimation problem, if the criterion function 18 is smooth then a derivative based method, such as Gauss Newton is appropriate. This may not be always the case. Recently Chernozhukov and Hong (2003) have offered an estimation approach that borrows from Bayesian estimation methods and is particularly suitable for complex problems and non-smooth criteria functions. Borrowing from the Bayesian literature they define the "quasi posterior" distribution of the parameters as

$$g_N(\theta|D) \propto \exp(-NQ(D|\theta))\pi(\theta) \tag{19}$$

where $\pi(\theta)$ is a suitable prior. Asymptotically the prior will not matter, but in any fixed sample the choice will affect the parameter estimates. The key result by Chernozhukov and Hong (2003) is that if we draw a sample of parameters $\theta^{(k)}$ from $g_N(\theta|D)$, where k denotes one random draw, then the sample mean of the $\{\theta^{(k)}, k = 1, K\}$ converges asymptotically to θ. To draw random vectors from $g_N(\theta|D)$ we can use Markov Chain Monte Carlo methods (Chib, 2001). This works as follows: guess an initial value for θ, say $\theta^{(0)}$, solve and simulate the model and compute the corresponding $\exp(-NQ(D|\theta))\pi(\theta)$, which is $g_N(\theta^{(0)}|D)$ up to an unknown constant of integration. Now define an update of θ by

$$\theta^{(k+1)} = \theta^{(k)} + \eta \tag{20}$$

where η is a random vector drawn from a distribution such that it respects any con-

straints on the parameter space. For example, for parameters that cover the entire real line η could be normal with some variance to be chosen by us and modified as the sampling proceeds until we reach a stationary distribution. Once the stationary distribution has been reached the sampling needs to remain the same.[15] We thus proceed as follows: compute the value of $g_N(\theta^{(k+1)}|D)$. The next element of the sample space that we will keep is

$$\begin{array}{ll} \theta^{(k+1)} & \text{with probability } p = \max(1, \frac{\exp(-NS(D|\theta^{(k+1)}))\pi(\theta^{(k+1)})}{\exp(-NS(D|\theta^{(k)}))\pi(\theta^{(k)}}) \\ \theta^{(k)} & \text{otherwise} \end{array} \tag{21}$$

Note that the constant of integration cancels out from 21 and hence never needs to be computed. This algorithm leads to a sample drown from $g_N(\theta|D)$. The estimator is then the average of the draws from the stationary distribution and the confidence intervals can be obtained directly from the quantiles of the sampled parameters. This approach can work well, even if the criterion function $Q(D|\theta)$ is not smooth because no derivatives are required.

Overall simulation methods require programming the solution of the model and our ability to simulate it but do not require the computation of an often intractable likelihood function. Such approaches promise to give a fresh impetus to the use of structural models. This is particularly important because as we shall see most methods that look simpler rely implicitly for their interpretation on particularly strong assumptions about individual behaviour.

[15] For implementation details of Markov Chain Monte Carlo methods see Robert and Casella (1999)

2.4 Estimating the Wage Returns to education in Mincer wage equations

Most of the work on the returns to education has not followed the approach described above; rather it has used the framework of the Mincer equations where educational attainment is summarised by years of education. Thus we digress to discuss the methods used that are based on years of education. The theoretical foundations for such a model can be found in the work of Mincer (1958, 1974). Heckman, Lochner and Todd (2003) have an excellent analysis of the theoretical foundations of the Mincer wage equation and on its empirical relevance.

The key differences in the underlying choice model that leads to the Mincer equation is the absence of direct costs of education and the absence of any uncertainty when education choice is made. Moreover, education is seen as enhancing human capital but not changing its nature: different education levels are perfectly substitutable for each other. The basic Mincer equation can be written as

$$\ln w_{it} = a_t + bs_i + cx_{it} + dx_{it}^2 + u_{it} \tag{22}$$

where s_i represents the years of schooling and x_{it} represents years of actual work experience. This relationship is derived in Mincer (1974) and results from pre-labour market investments in schooling and post school training. In the simplest form of the Mincer model, where there are no direct costs of schooling, but only an opportunity cost, and no heterogeneity in discount rates, the coefficient on schooling is the return to education and will be equal to the interest rate. However, in more complex models, where there are direct costs and possibly heterogeneity in discount rates and in costs, this is no longer true. So the first (well known) point is that in general just using a wage equation does not provide us with enough information to estimate the return to education, but just the educational premium for wages; this is just part of the story. In what

follows, we refer to the *wage return* to education as the effect of extra education on wages earned, rather than the full return, which would include a complete accounting of the costs and benefits.[16]

Relaxing a number of assumptions underlying the Mincer model we can end up with a relationship that is both nonlinear in education and where the returns differ across individuals. Perhaps the easiest way of justifying this equation is via the human capital production function, which can take a variety of functional forms and then price out human capital by the equilibrium in the labour market.

We can simplify the problem by replacing actual experience (the number of periods worked) and including potential experience ($Age_{it} - s_i$) in a linear fashion to start with. We will also abstract from all other relevant observable characteristics and we will drop the time subscript. However we will allow the baseline wage to be different across individuals by specifying that a depends on individual i (a_i) due to unobserved labour market ability. Thus we get

$$\ln w_i = a + bs_i + cAge + [a_i - a]$$

If differences in ability are known at the time the individual makes educational choices and are taken into account by them, the years of education will depend on them and be endogenous.[17] This means that $E[a_i - a|s_i] = q(s_i)$ where $q(s_i)$ is some non-trivial function of schooling s. In this simple framework, suppose we have a variable z_i which satisfies the rank condition $E(s_i|z_i) \neq 0$ (correlated with education) and the exclusion restriction $E(a_i - a|z_i) = 0$. Then the instrumental variable estimator is the sample analog of

$$b = \frac{E\left((z_i - \bar{z}) \ln w_i\right)}{E(s_i(z_i - \bar{z}))} \tag{23}$$

[16]The Mincer equation can also be viewed as a restrictive version of the Roy model, where the unobservables are the same across education groups.

[17]see Griliches (1977a,b) for one of the first and comprehensive discussions of the role of ability in wage equations.

which can be implemented by replacing the numerator with the sample covariance of z and log wages and the denominator by the sample covariance between z and years of schooling. The resulting estimator is consistent, i.e. converges to the true value b as the sample tends to infinity, under the stated conditions.

Of course the key difficulty is finding variables that somehow affect schooling and do not affect wages. Because of the linearity of the relationship our task is made easier: even a binary instrumental variable would be sufficient to estimate b. An obvious possibility is to use variables that reflect the cost of schooling when the individual was making these decisions. The key problem of course is that such cost related variables may also be related to the future productivity of the individual when they enter the labour market. A number of variables have been used in the literature in this respect. For example, Card (1995) discusses the use of distance from school as an instrument, which reflects both time and money costs of schooling. While this is clearly a cost related variable and is correlated with schooling it may also be correlated with individual ability. This is because individuals and schools are unlikely to be randomly allocated.[18] An example is individuals living in a city tend to be higher ability and probably more dynamic and ambitious. At the same time the increased population density will mean that schools are on average closer to individuals.

Other instruments relate to changes in legislation. Prominent examples include the use of changes in compulsory schooling laws by Harmon and Walker (1995). They exploit the fact that compulsory schooling increased twice in the UK, from 14 to 15 and then from 15 to 16. However, because their estimation involves comparing outcomes across successive cohorts, they are not able to allow for other confounding factors that may influence productivity and the returns to experience of each of the successive cohorts.

Meghir and Palme (2005) directly evaluate the impact of an educational reform that increases compulsory schooling and abolishes early streaming at 12 years of age.

[18] see also Card (2001) on this point.

In their case they are able to exploit the fact that the reform was introduced gradually across different municipalities in Sweden, which meant that at the same point in time there were municipalities operating different school systems and whose workers would eventually end up in the same labour market. This implied that they could compare across cohorts living in municipalities that switched education system between the cohorts to those living in municipalities that kept the old system for both cohorts (or indeed who were in the new system for both cohorts). The results showed clear benefits for individuals from lower socioeconomic groups both in terms of educational attainment and wages; for those from higher socioeconomic groups there was no effect of education but an adverse effect on their earnings over the lifecycle. Indeed the nature and scope of the reforms were such that we would expect them to affect wages directly as well as possibly through years of education. Hence in this case the reform does not provide a valid instrument for the returns to education, but can be evaluated as a policy in itself.[19]

This brings up an important point of general interest: it is often the case that reforms are used as instruments for estimating returns to education. However, if the reforms changed other aspects of education, such as its quality, it is no longer an excludable instrument for the quantity of education. The reforms to the number of compulsory years of schooling are a strong case at point: if we increase the number of compulsory years of schooling we may change both the peer group of those who would have continued anyway and possibly the pupil teacher ratio. And if we do increase the resources we will change the composition of the secondary teacher population, all of which can have a direct effect on wages. In most plausible cases the reform will not be excludable from the wage equation *a priori.*

Angrist and Krueger (1991) use the quarter of birth of an individual an instrument for education in a wage equation. Quarter of birth interacts with the laws on compuls-

[19]It did however, allow inferences to be made on the effects of streaming on groups whose socioeconomic background was such that their educational attainment could not be affected by the change in the compulsory schooling laws.

ory schooling to generate differences among individuals who happen to have been born on different dates: the reason this instrument may explain differences in the amount of schooling received is because one can drop out of school on their 16th birthday; hence depending on the month of birth some individuals have effectively fewer months of compulsory schooling than others. Interestingly this instrument acts at the individual level and does not affect aggregate schooling, as do reforms to the schooling laws mentioned above. Angrist and Krueger show that there are differences in total schooling by month of birth. However, the differences are small and this particular study led to the important literature on the effects of using instruments that are only weakly correlated with the variable to be instrumented (here schooling). One of the conclusions of this literature is that when an instrument is weak the estimated results are biased towards OLS (see Bound, Jaeger and Baker, 1995 and Steiger and Stock, 1997). A further issue with quarter of birth is whether it is excludable as an instrument from the wage equation: children born in different quarters start attending school at different ages, which may well have an impact on their performance. The effect may be small, but it has to be compared with the small effect that quarter of birth has on attained schooling.

In a interesting paper Acemoglu and Angrist (1999) combine compulsory schooling reforms as instruments that change both the individual level of education and the aggregate one with quarter of birth, that changes only individual decisions but have no aggregate effect. Their purpose is to distinguish between the private and the social returns to education; this brings to the fore the issues we discussed earlier: reforms to compulsory schooling laws affect directly those individuals who would have dropped out anyway. However, because they increase education for a number of people at the same time they change the composition and amount of educated individuals in the state. If education has externalities, i.e. social returns over and above the private ones, then using reforms as an instrument should pick up these effects; the estimated returns to education will be different from those measured by an instrument that "varied" indi-

vidual levels of education but not the aggregate. Indeed they will be larger if education has positive external effects. To measure the contribution of the external effects they then use quarter of birth instruments to identify the private returns as in Angrist and Krueger (1991). The difference of the two estimates should be the externality effect of education. However, if the reforms also affect the quality of education and equilibrium returns the estimates obtained with the reforms will be confounded by GE effects and quality differences. A further issue arises if returns to education are heterogeneous because, even under the monotonicity assumption the instruments may be measuring the returns corresponding to different types of individuals. Moreover, with GE and peer effects monotonicity may no longer be valid . In general it is particularly difficult to find satisfactory instruments for education returns without embedding the problem within the a structural model, which allows us to account for such confounding factors.

2.4.1 Nonparametric Models

When the wage returns to education are not constant but depend on education, as would be the case if the effect of schooling on wages were nonlinear, the instrumental variables approach becomes more demanding. Dropping age for notational simplicity, a general way of describing the problem is through the following model

$$\ln w_i = b(s_i) + u_i \tag{24}$$

where now $b(s_i)$ is some *unknown* function, and education may be endogenous. The econometric problem of using instrumental variables in this nonparametric context has been addressed by Newey and Powell (2003) and Darolles, Florens and Renault (2000). The estimator is defined based on the assumption that $E(u|Z) = 0$ as in the usual linear instrumental variables context. This restriction means that the error term u is mean independent of *any* nonlinear function of z. However, this is not sufficient: we also

need a suitable rank condition. Thus in this regard Newey and Powell (2003) introduce the critical rank condition that any function of education $\delta(s)$ can be predicted by Z. To restate their proposition 2.1:

[Proposition 2.1, Newey and Powell (2003) page 1567] If $E(u|Z) = 0$ *then* $b(s)$ *in 24 is identified if and only if for all* $\delta(s)$ *with finite expectation* $E(\delta(s)|Z) = 0$ *implies* $\delta(s) = 0$.

The importance of this result lies in its implications (or requirements) for identification. The practical difficulty is finding instruments that can satisfy these conditions: the rank condition is much more demanding than the equivalent one in a linear context because it requires that the instrument can produce predictions of any nonlinear function of education and that all these predictions are full rank. However, the practical importance of this theorem lies in what it tells us about the identifiability of such general nonlinear relationships.

Suppose we do have an instrument satisfying the conditions, we now briefly describe implementing an estimator for the function $b(s_i)$. To understand how the nonparametric estimator works we will start with the simple case where the education variable takes K distinct values (1 year to 22 years say) and we have an instrument Z which takes M distinct values; imagine this as reflecting discrete costs of schooling. The exclusion restriction effectively implies that the values of this instrument have been randomly allocated to individuals[20]. Since s_i takes K discrete values, $g(s_i)$ also takes K discrete values. The exclusion restriction implies that $(\ln w - b(S)|Z = z^j) = 0$.[21] This implies the following set of equations

$$E(\ln w|Z = z_j) = E(b(S)|Z = z_j) \qquad j = 1, ..., M \tag{25}$$

This represents a system of M equations with K unknowns. For example the first

[20]The theory requires the instrument just to be mean independent of the residuals, conditional on other observable characteristics. Hence the assumptions are weaker than complete randomisation that indices full independence.

[21]We use a capital to denote a random variable and a lower case to denote a specific realisation.

equation will have the form

$$E(\ln w|Z=z_1)=\Sigma_k b(S=s_k)\Pr(S=s_k|Z=z_1)$$

For this system of equations to have a unique solution for the K unknowns $b(S=s_k)$, $k=1,...,K$, we need the matrix whose (j,k) element is $\Pr(S=s_k|Z=z_j)$ to have rank K. This means that the instrument has to take at least K values and that the probability of different levels of schooling vary sufficiently with the instrument. This rank condition is the discrete analog of the Newey and Powell condition.

To implement this define the sample average $\log w$ when the instrument takes the value z^j by $\bar{y}^j$ whose value is determined by $\bar{y}^j=\frac{1}{N_j}\sum_{i=1}^{N}\left(1(Z_i=z^j)\ln w_i\right)$, where $1(Z_i=z^j)$ is one whenever in the sample the instrument take the value z^j and N_j is the number of such sample points. Denote by $p_{kj}=\Pr(S=s_k|Z=z_1)$ and by $\hat{p}_{kj}$ the sample estimate of this probability. The sample analog of 25 is the set of equations

$$\bar{y}^j=\sum_{k=1}^{K}b_k\hat{p}_{kj}\qquad j=1,...,M \tag{26}$$

where $b_k=b(S=s^k)$ is the set of k unknown values. Estimating g then involves simply solving the system of equations in 26. For the exact identification case ($M=K$) this simply means

$$\hat{g}=\hat{P}^{-1}\bar{y} \tag{27}$$

where $\hat{g}$ is the $K\times 1$ vector of all $\hat{g}_k s$ and $\hat{P}$ is the $K\times K$ matrix of probabilities of the $\hat{p}_{kj}$. For the overidentified case ($M>K$ and rank(P)=K) the estimator of g_k would minimise the distance

$$D(g)=(\bar{y}-Pg)'\Omega^{-1}(\bar{y}-Pg)$$

with respect to the vector g. In the above Ω is a suitable covariance matrix.

Extending this procedure to the continuous education case, involves solving an "ill-

posed inverse" problem, where in general the matrix $\hat{P}$ in 27 is not invertible in finite samples. The approach to solving this is called regularisation, which involves adding a component to P so that it becomes invertible, for example replacing P by $P^* = P + \lambda_N I$, λ_N being a scalar that declines (at a suitable rate) as the sample size N goes to infinity and I being the identity matrix. Darolles, Florens and Renault (2002) and Newey and Powell (2003) offer solutions to the estimation problem in more general terms than we described here.

The key implication of this discussion is that a wage equation that is linear or nonlinear in years of education can in principle be estimated by instrumental variables, without saying much about the structure of the education choice model, other than the standard conditions on the instruments. The most important restriction that has been imposed in this discussion is that the returns to education are homogeneous, or more precisely that education choice does not depend on heterogeneous returns to education.

2.4.2 Heterogeneous returns to years of education and nonparametric models

If individuals have say different learning abilities the wage returns to education may differ across individuals. A simple way of expressing this is to rewrite the Mincer model as

$$\ln w_i = a_i + b_i s_i + cAge. \tag{28}$$

where a_i and b_i are unobservables. Now rewrite the above in the form

$$\ln w_i = a + bs_i + cAge + [a_i - a + (b_i - b)s_i]$$

where the term in square brackets $[a_i - a + (b_i - b)s_i]$ is the residual . If the individual takes account of a_i and b_i in choosing s_i, or indeed if any mechanism allocating schooling to individuals depends on a_i and b_i then OLS will be inconsistent for $b = E(b_i)$. So

the question is how can we estimate b or other interesting features of the *distribution* of the wage returns. These may include the impact of education level s for those who choose that level (analogous to the effect of treatment on the treated). When s_i is binary, this is the subject of the extensive treatment effects literature, which has also been extended to continuous treatments.[22] We discuss the binary or multiple discrete case in the following section, when we go back to our theoretical framework of education choice and wages.

To start off we define the reduced form model for education choice. Thus we specify

$$S = P(Z) + V \tag{29}$$

where we define $P(Z) \equiv E(S|Z)$.

The first important lesson from this literature is that Instrumental Variables as defined in 23 is not consistent for the average parameter $b = E(b_i)$, i.e. for the average returns to education, without further restrictions. Heckman and Vytlacil (1998) show that IV is consistent for $E(b_i)$ with the additional assumption that $E((b_i - b)v_i|Z) = 0$, implied by the stronger assumption $(A,B,V) \perp\!\!\!\perp Z$, where the capital letters are the random variables and a_i, b_i and v_i are their specific realisation respectively.

To see that this is a strong assumption suppose that the instrument Z was randomised. In the standard IV framework, this would be sufficient for identification, because randomisation guarantees that $E(a_i|Z) = 0$. Now however, we need to take a stance about the actual model generating educational choices. Suppose for instance that the true model generating educational choices took the form $S = D(Z,u)$ where u is unobserved heterogeneity. While $(a_i,b_i,u_i) \perp\!\!\!\perp Z$ is guaranteed by randomisation we require the stronger assumption that $(a_i,b_i,v_i) \perp\!\!\!\perp Z$ for $v = S - E[D(Z,u)|Z]$ or at that $E[b_i v_i|Z] = 0$; this does not follow without further assumptions. The implication of this

[22] see Heckman and Robb (1985), Heckman, LaLonde and Smith (1999), Imbens and Angrist (1994), Florens, Heckman, Meghir and Vytlacil (2008) , Imbens and Newey (2009) and Altonji and Matzkin (2005) amongst many others.

discussion is clear: even if the assumptions underlying IV are valid, the interpretation of IV coefficients is unclear and will depend on the structure of the education model itself. We now go deeper into this issue.

More generally, suppose the education model is nonlinear in education s_i so that we can write (ignoring other variables)

$$\ln w = g(S, e) \tag{30}$$

where e is a vector of unobserved characteristics and where S will in general depend on e. A number of papers have attempted to tackle this important problem in various ways, by making different assumptions and considering identification of different aspects of the model. These include Chesher (2003), Altonji and Matzkin (2005), Imbens and Newey (2009) and Florens, Heckman, Meghir and Vytlacil (2008).[23] Without getting into too much detail we briefly review some of these here.

Altonji and Matzkin develop two approaches. In the one that is most relevant to our problem they make a conditional independence assumption that the distribution of the error term $f(\cdot)$ is such that $f(e|Z,S) = f(e|Z)$, i.e. conditional on Z the distribution of the error term does not depend on schooling. In this case e can be a two dimensional vector of errors that affects log wages in some arbitrary way. The authors identify the average effect of schooling on wages at each level of schooling s, i.e. $E(\frac{\partial g(S,e)}{\partial S}|S = s)$, based on their assumption.

Chesher (2003) develops identification results for the impact of the endogenous variable (here schooling) on quantiles of the distribution of the outcome. His identification results rely on weaker than usual *local* independence conditions; these require that specific quantiles of the distributions of the unobservables are insensitive to changes in the instrument. He also requires that the outcome of interest, here the wage, is monotonically related to the unobservables. In this his model is more restrictive than that of

[23] Blundell and Powell (2004) show identification and estimation for nonseparable models with a binary dependent variable.

Florens et al. (2008) and Imbens and Newey. If all quantiles of the unobservables are insensitive to the instrument, then global identification follows.

One issue that is important is that the local independence conditions do not have a clear relationship with an underlying choice model. While we can specify sufficient conditions on behaviour for full independence to be satisfied no such conditions have been specified for local independence, when full independence is not valid.

Imbens and Newey (2009) consider a general case where e in equation 30 is a vector of unobservables, and hence they allow for a completely flexible specification of heterogeneity in 30. To prove identification they specify the equation assigning values to the endogenous variable, which in our case is the model of education choice to take the form

$$S = P(Z,U) \tag{31}$$

where the function P is *strictly monotonic* in U. They define the control variate $V = F_{S|Z}(S,Z)$; in a binary choice context this would be the probability of $S = 1$ give Z.

The core of the identification result in their paper is based on the following three assumptions:

1. The function P in 31 is *strictly monotonic* in U;
2. The errors (U,e) are independent of Z,i.e. $(U,e) \perp\!\!\!\perp Z$. This implies that S and e are independent conditional on V;
3. The support of V given S is the same as the support of S.

To understand the meaning of this last assumption return to the definition of V and suppose Z does not affect V for some $S = s'$; in this case the support of V given $S = s'$ would be degenerate. Thus this assumption requires Z to affect V, which makes it equivalent to a rank condition. However, it also has another important implication

because it requires Z to be able to span the entire support of V whatever schooling level we consider. To see why this may be restrictive, suppose U is unobserved ability and Z are the observable costs. The assumption effectively requires that Z varies in such a way as to ensure that all ability levels (U) are represented within each schooling level S.

Under these assumptions the authors prove identification of the "quantile structural function", i.e. they can identify the quantiles of $g(S,e)$, defined as $q(\tau,S)$ where τ stands for the quantile of $g(S,e)$.This allows them to identify quantile effects, i.e. how changes in education change the τth quantile of wages ($q(\tau,S=s)-q(\tau,S=s')$). They also derive a number of other identification results under weaker assumptions, which are beyond the scope of this chapter.

Florens, Heckman, Meghir and Vytlacil (2008) consider a more restrictive class of models, but obtain identification under a weaker rank condition, namely that of measurable separability, discussed below. The class of models they consider take the following nonseparable form

$$\ln w = m(S,\varepsilon) \equiv g(S) + \varepsilon_0 + \sum_{k=1}^{K} S^k \varepsilon_k \tag{32}$$

where the function $g(\cdot)$ is not known. The model discussed in Newey and Powell (2003) and Newey, Powell and Vella (1999) is one where $K=0$. More generally, this function is non-separable in education S and unobserved heterogeneity. The restrictions *vis-a-vis* Imbens and Newey (2009) is that $g(S)$ has to be differentiable up to the order K and the maximum value of heterogeneous terms K needs to be known in advance. In the context of Florens et al. (2008) , no identification results have been proved for the case where K is unknown and has to be estimated. The object of interest for the Florens et al. (2008) core identification result is the "Average Treatment Effect", or the

average wage return to education at each level of education S. This is defined as

$$\Delta^{ATE}(s) = \frac{\partial g(s)}{\partial s} = E_{\varepsilon}\left[\frac{\partial m(S,\varepsilon)}{\partial S}\right] \tag{33}$$

The econometric problem is to identify $\frac{\partial g(s)}{\partial s}$ or even the function $g(S)$. As stated above, standard instrumental variables will not work. Florens et al. (2008) make the following assumptions (omitting some technical details).

1. The function g is differentiable to the Kth order.
2. Control Function: $E(\varepsilon_j|Z,S) = r_j(V)$ for $j = 0,...,K$, where $r_j(\cdot)$ is a known or identifiable function and V is defined in 29.
3. Rank condition (measurable separability): S and V are measurably separated, that is, any function of S almost surely equal to a function of V must be almost surely equal to a constant.

Theorem [from Florens et al. (2008) , Theorem 1 p 1197] Given assumptions 1, 2 *and* 3 *above the average wage returns to education in 33 as well as the wage returns to education for those who chose schooling level* $S = s$ *are identified.*

This identification theorem defines conditions under which we can actually identify the wage returns to education even with errors that enter in a non-separable way and in a very general fashion. It is important of course to understand the limitations implied by the assumptions.

Assumption 1 precludes any kinks or discontinuities in the relationship between education and wages; discontinuities could be induced by sheepskin effects, where wages may jump discontinuously upon graduation for instance.

Assumption 2 is similar (but not identical) to the usual exclusion restriction. It states that all the dependence between the unobservables in the education equation and the educational assignment rule can be expressed through some known or identifiable

function of the residual in the educational equation reduced form. This assumption is the same as the one used in Heckman (1979) and many others since;[24] In Heckman (1979) the control function is the Mills ratio. Imbens and Newey (2009) use the same concept of a control variate. This control variate induces conditional independence between schooling and unobserved heterogeneity. The third assumption is a rank condition: it requires sufficient independent variation of V and S; If the instrument Z did not "explain" S this condition would not be satisfied. Note however, that measurable separability does not require the support of V given S to be the same as the support of V. As such it is a weaker assumption than the one used by Imbens and Newey (2009).

These assumptions seem overly technical; we need to explain what they mean in terms of economic behaviour. Unfortunately necessary and sufficient conditions are not available. Florens et al. (2008) provide sufficient conditions on a structural model of education choice for the conditions to be satisfied. In particular they posit a model where the education choice can be written as 31 where U is a continuous scalar random variable and P is an increasing function of U. This is restrictive, because it requires that just one unobservable factor characterises educational choice and that this variable is monotonically related to education. For example, if the level of education depends on unobservable costs and on labour market ability then S will depend on two unobserved factors that may not be possible to aggregate them into one satisfying the monotonicity assumption. Florens et al. (2008) show that if the instrument Z is independent of all unobservables in the model i.e. $Z \perp\!\!\!\perp (U, \varepsilon_0, \varepsilon_1, ..., \varepsilon_K)$, then under the assumption of measurable separability between U and S assumptions 2 and 3 of the theorem are satisfied and identification of the average wage return to education follows. Hence results obtained under the control function assumption can be interpreted in the context of any education choice model that can be expressed as $S = S(Z,U)$, which is monotonic in U.

[24] See Heckman and Robb (1985) and Newey, Powell and Vella (1999)

2.4.3 Education choice and Wages: A Simple Illustration and Discussion

Florens, Heckman, Meghir and Vytlacil (2008) present the following example to illustrate the issues. Suppose that the discounted annualised earnings flows for s years of education is W_s .

$$W_s = \varphi_0 + (\varphi_1 + \varepsilon_1)s + \frac{1}{2}\varphi s^2 + \varepsilon_0$$

and the cost function for schooling as

$$C_d = C_0(Z) + (C_1(Z) + v_1)s + \frac{1}{2}C_2(Z)s^2 + v_0 \tag{34}$$

where ε_k and v_k $(k = 0, 1)$ are, respectively, unobserved heterogeneity in the wage level and in the cost of schooling. We impose the normalisations that $E(\varepsilon_k) = 0$, $E(v_k) = 0$, for $k = 0, 1$. We implicitly condition on variables such as human capital characteristics that affect both wages and the costs of schooling. The Z are factors that only affect the cost of schooling, such as tuition costs.

Assume that agents choose their level of education to maximise wages minus costs. Let S denote the resulting optimal choice of education. S solves the first order condition

$$(\varphi_1 - C_1(Z)) + (\varphi_2 - C_2(Z))S + \varepsilon_1 - v_1 = 0.$$

Assuming that $\varphi_2 - C_2(Z) < 0$ for all Z, the second order condition for a maximum will be satisfied. This leads to an education choice equation (assignment to treatment intensity rule)

$$S = \frac{\varphi_1 - C_1(Z) + \varepsilon_1 - v_1}{C_2(Z) - \varphi_2}.$$

This choice equation satisfies the monotonicity restriction discussed above and if Z is randomised it will be jointly independent of ε_k and v_k $(k = 0, 1)$. This implies the control function assumption and the model is identified without knowledge of or further restrictions on the functional form of the wage and the education choice equations.

However this result is sensitive to changes in both the degree of heterogeneity in wages or directly in the cost function because both can affect the structure of educational choice. Consider the same example as before, except now the second derivative of W_d is also stochastic:

$$W_s = \varphi_0 + (\varphi_1 + \varepsilon_1)s + \frac{1}{2}(\varphi_2 + \varepsilon_2)s^2 + \varepsilon_0$$

In itself this poses no problem, except that in an optimising model it will change the structure of educational choice; this now becomes

$$S = \frac{\varphi_1 - C_1(Z) + \varepsilon_1 - v_1}{C_2(Z) - \varphi_2 - \varepsilon_2}$$

In this case, the structural model makes S a function of $V = (\varepsilon_1 - v_1, \varepsilon_2)$, which can satisfy the independence assumption if Z is randomised, i.e. $Z \perp\!\!\!\perp (V, \varepsilon_0, \varepsilon_1, \varepsilon_2)$ *but* V is not a scalar error. We can still construct an education model depending monotonically on one error term but this new error term will not generally be independent of Z: define a residual $\tilde{V} = F_{S|Z}(S|Z)$; the "reduced form" education choice equation can then be written as $S = \tilde{g}(Z, \tilde{V}) = F^{-1}_{S|Z}(\tilde{V}|Z)$, which is increasing in $\tilde{V}$. Thus S is strictly increasing in a scalar error term $\tilde{V}$ that is independent of Z by construction. However, Z is not independent of $(\tilde{V}, \varepsilon_0, \varepsilon_1, \varepsilon_2)$ as required by the identification theorem, despite the fact that it is independent of the original errors $(V, \varepsilon_0, \varepsilon_1, \varepsilon_2)$. To see why, note that

$$\Pr(\tilde{V} \leq v | Z, \varepsilon_0, \varepsilon_1, \varepsilon_2) =$$

$$\Pr\left[v_1 : \tfrac{\varphi_1 - C_1(Z) + \varepsilon_1 - v_1}{C_2(Z) - \varphi_2 - \varepsilon_2} \leq F^{-1}_{S|Z}(v) \middle| Z, \varepsilon_0, \varepsilon_1, \varepsilon_2\right] \neq \Pr(\tilde{V} \leq v | \varepsilon_0, \varepsilon_1, \varepsilon_2)$$

In the above varying Z changes the set of v_1 for which the condition is true and hence the distribution of $\tilde{V}$ depends on Z.

This example illustrates how the specification of the educational choice model has

implications for the identifiability of the wage returns to education. As emphasised by Heckman, LaLonde and Smith (1999) and Abbring and Heckman (2007) informational assumptions can play an important role: here if the individual knows and takes into account the complete structure of wages the educational choice model becomes such that the sufficient assumptions for identification used by most papers in this literature may no longer be valid depending on the degree of heterogeneity in the wage equation. Hence even in a non-parametric framework and with randomised instruments the interpretation of results will depend crucially on the model driving educational choice: randomisation of the instruments is not sufficient in this respect. Thus, what transpires from the above is that identification depends on the nature of education choice, beyond the simple statement that a valid instrument is available.[25] This contrasts to an extent with what is known about models with homogeneous effects.

In the next section we will consider explicitly educational choice as discrete. This framework permits consideration of identification and estimation issues based on richer models of educational choice.

2.5 Identification and Estimation of the wage returns to Education in the dynamic discrete education choice model.

We now return to the dynamic multisector model of education choice described in section 2.2. This is a more complex model because it recognises the sequential nature of education choice and allows for uncertainty, which gets revealed gradually between different educational stages. It also allows for the possibility of comparative advantage for a particular educational level.[26] Our aim is to review approaches to identifying and estimating measures of the returns to education, such as the average treatment effect or the Local Average treatment effect.

[25] valid in the traditional sense of being uncorrelated to unobservables and correlated with education.
[26] See for example Heckman and Sedlacek (1985).

There is a vast literature on discrete treatment effects and their identification. Some of the most important results are presented in Heckman and Robb (1985), Heckman, LaLonde and Smith (1999), Imbens and Angrist (1994), Heckman and Vytlacil (2005), Carneiro, Heckman and Vytlacil (2010) and many others. Most papers define statistical assumptions that lead to identification. Some make the additional important step of relating these assumptions to the underlying economic behaviour a prime example being Vytlacil (2002).

We focus on two issues that arise when using the model of section 2.2 as an organising framework. There are two issues. First to what extent is the full model nonparametrically identified and second if the full model is not identified under what conditions can we at least identify the marginal distribution of earnings for each education level, so as to get to the average returns to education.

Magnac and Thesmar (2002) explicitly analyse the identification in dynamic discrete choice models with uncertainty where the data only include the discrete choices as well as observations on the relevant state variables. In our case these would be the education choices and the variables determining the costs of education respectively (Z). No observations on outcomes motivating such choices, such as income are observed. They show that even in the absence of persistent unobserved heterogeneity the model is seriously underidentified: to identify the within period utility, without functional form restrictions one needs to know the distribution of the shocks, the discount factor and the current and future preferences for a reference alternative. Exclusion restrictions between alternatives can improve things but not by much.

However, when we observe outcome variables such as earnings and when we can link the choice of a level of education to the observed outcome as indeed the model presented in section 2.2 the prospects for identification improve. Heckman and Navarro (2007) present a number of identification theorems relating to the dynamic structural model itself and to the distribution of earnings in each education level. Many of the

issues can be understood by taking the simpler framework of Heckman, Urzua and Vytlacil (2006a,b) (HUV henceforth).

Consider first a simple framework where education choice can be expressed as a once and for all choice at a point in time. Individuals choose the level of education among a set of possible levels. However the levels are not necessarily ordered. The underlying reason why the choices are not ordered are the dynamics: it is possible that an individual choosing between dropping out of school or attending high school, could choose the former, in the absence of any other choice, but that if the choice of college is added, then they could progress to college (via high school graduation). HUV study identification of models with discrete treatments and unordered choices and provide identification results, which we outline briefly here.

Write the net payoff to education as

$$R^J(Z^J,X) = \vartheta^J(Z^J,X) - V^J,\ J = S,H,C \tag{35}$$

where Z^J is the set of variables that affect education choice J; these could be the costs that affect a particular education level, such as fees or transport costs to the closest educational institution. Because of dynamics all Z^J may be the same (see below). Let $Z = \{Z^S, Z^H, Z^C\}$. The payoff to education is earnings and is given by 5. Now make the following assumptions as in HUV

1. The unobservables are jointly independent of Z, X and *age*: $(\tau^S + \varepsilon_t^J, V^J,\ J = S,H,C) \perp\!\!\!\perp \{Z,\ age,\ X_i\}$

2. The support (supp) of the functions $\vartheta^J(Z^J)$ and $m^J(age,X)$ is independent of

each other so that

$$supp\{\vartheta^J(Z^J,X), m^J(age,X), J = S,H,C\} =$$

$$\vartheta^S(Z^S,X) \times m^S(age,X) \times \vartheta^H(Z^H,X) \times m^H(age,X) \times \vartheta^C(Z^C,X) \times m^C(age,X)$$

3. The structures of the functions $\vartheta^J(Z^J)$ and of the variables Z^J is such that their support is at least as large as the support of V^J :

$$\text{supp}\{\vartheta^J(Z^J)\} \supseteq \text{supp}\{V^J\}$$

4. Given age and Z, X has full rank

These assumptions imply that we can find combinations of values of Z such that the probability of any choice J becomes 1. Within that "limit set" as Heckman and Navarro (2007) call them we can identify the marginal distribution of earnings Y^J conditional on age and X. The latter follows from the independence assumption that ensures that the distribution of earnings is the same for whatever value of Z and by the rank condition that ensures that whatever the value of Z and age there is sufficient variation in X. In addition if all we are interested is average earnings given X and age, then all we need is that the errors in the earnings equation are mean independent of Z, *age* and X.

This identification result suggests an estimation strategy for mean earnings. Suppose the only X regressor was age. Then we can estimate mean earnings for education level J at age a as

$$\hat{Y}(a,J) = \frac{\sum_{i=1}^{N} K(\hat{p}(Z_i) - 1) Y_i^J (age_i = a)}{\sum_{i=1}^{N} K(\hat{p}(Z_i) - 1)}$$

where $K(\hat{p}(Z_i) - 1)$ is a kernel giving maximum weight when $\hat{p}(Z_i) = 1$. This is a

weighted average of the earnings of (potentially) all individuals with education level J, with the weights being higher the higher the predicted probability of attaining that level. The weights towards individuals with probability equal to one of achieving this level increase as the sample size increases, but at a rate which is slower than the sample size increase. Clearly this estimation procedure is only justified if the limit sets exist in the population and the assumptions detailed above are justified. Below we discuss further the support assumptions and the consequences of them being violated. But before this we turn to the model that is explicitly dynamic.

The question is how different is the dynamic context in terms of the required assumptions for identifying the marginal distributions of earnings corresponding to different education levels. By examining equations 12, 13 and 14 it is apparent that all probabilities depend on all Zs so long as these are all known when the decisions are made sequentially over time. Second it is also apparent that the probabilities are non-linear functions of unobserved heterogeneity and the decision problem is not separable in observables and unobservables as in 35. The education choice model has the non-separable form

$$R^J = \vartheta^J(Z,X,e)\ J = S,H,C \tag{36}$$

where R^J is the lifecycle value of alternative J and where e is a vector of unobservables. Of course there is some structure to this, which will matter both in terms of understanding whether the assumptions are valid or not and for identifying the dynamic discrete choice model. Heckman and Navarro (2007) make assumptions on the primitives of the model so that the support conditions discussed above carry over to this context. As before we need to be able to argue that the required limit sets exist. Given they exist the same identification argument applies as before.

2.6 Using Bounds to estimate the returns to education

The idea of limit sets is interesting intellectually but in practice it is highly unlikely that we observe suitable variables such that the limit sets exist. The variables we observe are likely to have limited support; for example it is unlikely that we will observe such a range of fees that at one end all attend college (presumably this would require a hefty subsidy) and at the other no one did. Even combining this with other cost type variables like distance from college we are not likely to find a set where the probabilities of attendance reach the limit or even come close; the condition that the distribution of earnings conditional on other observables is the same within such limit sets, is even less likely to be satisfied. However, such identification strategies help us understand the nature of the problem and can sometimes points to other strategies. One strategy is to use the model we described in 2.2 or other such suitable specification, imposing distributional and other functional form assumptions. Indeed, even if the model is conceptually identified this may be a way of improving efficiency in practice. A diametrically opposite alternative is to make minimal assumptions and follow the route of partial identification and use bounds as discussed in Manski (1994), Manski and Pepper (2000) and Blundell, Gosling, Ichimura and Meghir (2007). In the context of this discussion the latter approach is particularly useful: it offers an alternative approach to learning something about the distribution of earnings with minimal assumptions and illustrates clearly the identification problem and how the limit set assumption resolves it.

Suppose we wish to identify the marginal distribution of earnings given a level of education J for individuals with characteristics X, $F(Y^J|X)$. Based on the law of iterated expectations we can write

$$F(Y^{J=j} < y|X) = F(Y^J < y|X, J = j)\Pr(J = j|X) + F(Y^J < y|X, J \neq j)\Pr(J \neq j|X)$$

where where $F(Y^J < y|X, J \neq j)$ is the distribution of earnings corresponding to level J for those that did not choose level J. This is not observed and without further assumptions all we can say is that lies in the closed interval [0,1]. Then this implies the "worst case" bounds

$$F(Y^J < y|X, J = j)\Pr(J = j|X)$$

$$\leq F(Y^{J=j} < y|X) \leq$$

$$F(Y^J < y|X, J = j)\Pr(J = j|X) + \Pr(J \neq j|X)$$

Because of the lack of any exclusion restrictions the lower and upper bounds can never be equal; in other words without further assumptions the distribution of earnings in each education group are never identified and of course neither is any notion of a wage return to education.

Now suppose we do possess a set of instruments Z as above from which education in level J is independent of. The assumption is that $F(Y^{J=j} < y|Z, X) = F(Y^{J=j} < y|X)$. Thus Z is excluded from earnings. However, these affect both the probability of selecting education choice J and through selection they also affect the conditional distributions. Thus for each value of the instruments Z and each education level J we have that

$$F(Y^J < y|Z, X, J = j)\Pr(J = j|Z, X)$$

$$\leq F(Y^{J=j} < y|X) \leq$$

$$F(Y^J < y|Z, X, J = j)\Pr(J = j|Z, X) + \Pr(J \neq j|Z, X)$$

Since the distribution of earnings does not depend on Z we can choose the best bounds

across Zs. Thus the tightest bounds are

$$\max_Z \left\{ F(Y^J < y|Z,X,J=j) \Pr(J=j|Z,X) \right\}$$

$$\leq F(Y^{J=j} < y|X) \leq \tag{37}$$

$$\min_Z \left\{ F(Y^J < y|Z,X,J=j) \Pr(J=j|Z,X) + \Pr(J \neq j|Z,X) \right\}$$

This links the approach with the earlier discussion: given the exclusion restrictions and assuming the distribution of X is not degenerate given Z, a necessary and sufficient condition for the distribution of earnings to be identified is that the upper and lower bounds in 37 to be equal. The existence of limit sets as discussed in Heckman and Navarro (2007) would imply such an equality. However this is just a sufficient condition, and it is possible that the bounds are equal without the limit set assumption. The difficulty is understanding how such conditions may relate to an underlying model of choice and can we expect them to hold.

If identification is not obtainable, i.e. if

$$\max_Z \left\{ F(Y^J < y|Z,X,J=j) \Pr(J=j|Z,X) \right\} < \tag{38}$$

$$\min_Z \left\{ F(Y^J < y|Z,X,J=j) \Pr(J=j|Z,X) + \Pr(J \neq j|Z,X) \right\}$$

for some are all X, then we can only obtain bounds on the distribution of earnings and thus bounds on its quantiles. Moreover, without bounds on the support of Y^J we cannot obtain bounds on means and variances and other moments, other than the order statistics. In this case we need to either impose restrictions on the support or compare the order statistics as a measure of the wage returns to education. To see how this works suppose we wish to compare quantile q. Define $w^{ql}(J=j,Z,X)$ the w that solves the

equation

$$q = F(w|Z,X,J=j)\Pr(J=j|Z,X) \tag{39}$$

and $w^{qu}(Z,X,J=j)$ the solution to

$$q = F(w|Z,X,J=j)\Pr(J=j|Z,X) + (1-\Pr(J=j|Z,X)) \tag{40}$$

Thus the upper bound of the qth quantile for a particular value of Z is equal to the $q/\Pr(J=j|Z,X)$ quantile of the observed distribution of earnings for those with education level j. The lower bound is the $(q-(1-\Pr(J=j|Z,X)))/\Pr(J=j|Z,X)$ quantile of the same observed distribution. To then use the exclusion restriction the best bounds can be obtained by

$$\max_Z \left\{ w^{ql}(Z,X,J=j) \right\} \leq w^q(X,J=j) \leq \min_Z \{ w^{qu}(Z,X,J=j) \}$$

Now suppose we define the wage returns to education as the difference in the medians of the two distributions. This measure of gain is bounded by

$$w^{ql}(X,J=C) - w^{qu}(X,J=H) \leq \Delta^{C/H} \leq w^{qu}(X,J=C) - w^{ql}(X,J=H) \tag{41}$$

where $q = 0.5$. This is not equivalent to the wage returns usually presented, which corresponds to comparing means. Thus, lack of point identification has led us to compare different aspects of the distribution of earnings across education groups.

To implement the bounds approach we can directly estimate the conditional distribution of earnings given each education group. BGIM (in the context of selection into work rather than education - see below) use

$$\hat{F}_N(w|J=j,age=a,Z=z) = \frac{\sum_{i=1}^N \Phi\left(\frac{w-w_i}{h}\right) I(J=j)\kappa_{az}(age_i,z_i)}{\sum_{i=1}^N I(J=j)\kappa_a(age_i,z_i)}$$

In the above $\hat{F}_N(\cdot)$ denotes the estimated distribution and $\Phi\left(\frac{w-w_i}{h}\right)$ is the standard normal distribution function and it is used instead of the indicator function $1(w_i \leq w)$ to provide some smoothness. As the sample size gets bigger we can reduce h, then this function becomes zero very fast as observations above w are used and is one for values of w_i even slightly lower than w. For the sample size in the BGIM study h was set at a fifth of the standard deviation of wages. For the function $\kappa_{az}(age_i, z_i)$ BGIM use

$$\kappa_a(age_i, z_i) = \left(\frac{age_i - a}{3} + 1\right)^2 \left(\frac{age_i - a}{3} - 1\right)^2 I(|age_i - a| < 3)\phi_k(z_i)$$

where

$$\phi_k(z_i) = \left(\frac{z_i - z_k}{0.2} + 1\right)^2 \left(\frac{z_i - z_k}{0.2} - 1\right)^2 I(|z_i - z_k| \leq 0.2).$$

Once the distributions of wages and the probability of attainment conditional on Z have been estimated we can then apply 38 to bound the distribution, which can then be used to estimate the bounds to the quantiles. Here an interesting observation can be made: the bounds that depend on the exclusion restrictions Z may cross. This can happen either because the restrictions are wrong or because the ample is small. We can thus devise a test of the null hypothesis that the bounds are equal against the alternative that the lower bound is above the upper bound. Rejecting implies the restrictions are invalid. The test has power against the alternative, but cannot detect invalid restrictions that do not lead to the bounds crossing. BGIM discuss such tests and implement it using the bootstrap. Kitagawa (2010) derives formally a test for independence in a similar context.

2.7 A special case: binary educational choice

Many problems in the broad area of education and training can be represented as the impact of a binary treatment, whose choice may be endogenous. A prime example

is the impact of vocational training. In this case the wage equation would take the form

$$\ln w_i = a + b_i T_i + u_i \tag{42}$$

or in counterfactual notation used in the treatment effects literature

$$\begin{aligned} \ln w_i^0 &= a + u_i \\ \ln w_i^1 &= a + b_i + u_i \end{aligned} \tag{43}$$

where w_i^0 represents wages in the no training state and w_i^1 represents wages in the training state. Estimation and identification of this model has been widely analysed with some of the key results to be found in Heckman, LaLonde and Smith (1999). We do not reproduce these here; however we complete our discussion of instrumental variables by pointing out the interpretation of this estimator in this context.

Suppose we possess a binary instrument Z; for example an indicator as to whether some individuals are facing a different policy environment, such as a training subsidy and which we take as having been randomly allocated. Suppose the model is the simple Roy model with just a binary education/training choice ($T_i = \{1,0\}$).

Instrumental variables for this problem has been analysed by Imbens and Angrist (1994). Define $T(1)$ and $T(0)$ to be indicators of whether an individual would take up training when $Z = 1$ and $Z = 0$ respectively. They assume that the instrument Z is jointly independent from all unobservables i.e. $(\ln w^0, \ln w^1, T(1), T(0) \perp\!\!\!\perp Z)$; they also make a critical monotonicity assumption: no person who would have obtained training when facing environment $Z = 0$ refrains from training when $Z = 1$. In terms of the notation above we have that for all individuals $T(1) \geq T(0)$.[27] Under these assumption they show that using Z as an instrument in the above regression (42) will

[27] Obviously all that matters is that Z_i induces either no movement or movement in the same direction for all individuals. So we need to have either that $T(1) \geq T(0)$ for *all* individuals or alternatively that $T(0) \geq T(1)$ for *all* individuals.

identify the effect of T on wages for those choosing $T = 1$ when facing $Z = 1$ (e.g. when offered the subsidy) and who would have chosen $T = 0$ if instead they faced $Z = 0$ (no subsidy).[28] The effect is known as the Local Average Treatment Effect (LATE). LATE is not invariant to the choice of instrument: different policies that act on different margins, i.e. induce different types of individuals into training, can lead to different estimates of the effect if it is heterogeneous. However, in each case the effect can be interpreted as causal. What is under question is the external validity/generalisability of the estimate. We are left with a situation where we can estimate the impact of training in some specific context, but we cannot necessarily generalise to other contexts.

Vytlacil (2002) shows that the LATE assumptions are equivalent to those of the traditional selection model. Thus the LATE assumptions are satisfied if and only if the training choice can be represented by a threshold crossing model, i.e.

$$T = 1 \Longleftrightarrow \varsigma > g(Z) \tag{44}$$

where the unobservables in 44 and 42 satisfy the independence assumption $(\varsigma, b, u) \perp\!\!\!\perp Z$. This means that if we can transform the decision rule implied by an economic model into the form 44 with all unobservables jointly independent of Z then that economic model is consistent with the assumptions implied by LATE.

Important cases where these conditions may not be satisfied is when there are general equilibrium effects or peer/congestion effects. Suppose that Z represents a subsidy to college education and suppose that many individuals take up college education. This may well discourage individuals with say high lnw_i^0 from attending college, when they may have done so without the subsidy.

Heckman and Vytlacil (2005) provide an elegant way of interpreting LATE and placing it in a broader family of treatment effects.[29] Suppose we represent binary

[28] i.e. the effect for those for whom $T(1) = 1$ and $T(0) = 0$

[29] see also Carneiro, Heckman and Vytlacil (2010)

education choice by a threshold crossing model, which satisfies the monotonicity restriction (Vytlacil, 2002)

$$\Pr(T_i = 1|Z) = \Pr(\kappa(Z) < v)$$

An individual is defined as marginal with respect to this training choice if $\kappa(Z) = v$: given Z the unobservable characteristics are such that the benefits and costs of training exactly outweigh each other. A small increase of the benefits will draw this person in. The marginal treatment effect at some value of Z is the effect of training on individuals who are just indifferent between accepting training and not, i.e. $b^{MTE}(z) = E(b_i|\kappa(Z = z) = v$. For convenience we can rewrite this relationship by defining $\kappa^u(Z) = F(\kappa(Z))$ so that we can define the marginal individual as $\kappa^u(Z = z) = p$ where $p \in [0,1]$. Now consider a policy that increases Z from a to b. Then the LATE parameter is the average effect for all those individuals in the range $[\kappa^u(Z = a), \kappa^u(Z = b)]$. If on the other hand the policy increases the value of Z by some fixed value, and assuming the LATE assumptions are valid at all Z, then LATE will be an average across different LATE values corresponding to the different starting points for Z. More generally, with Z being continuous one can imagine estimating an MTE parameter at all levels of Z. Heckman and Vytlacil (2005) show that the average MTE is the average treatment effect. Moreover, all treatment effect parameters, such as the average effect of treatment on the treated, can be expressed as weighted averages of the MTE over different relevant ranges. From a policy perspective the MTE offers us a way of estimating the marginal benefit in terms of our outcome variable (such as wages) of a small increase in say the incentive to obtain training.

Heckman and Vytlacil (2005) and Carneiro et al. (2010) show that that the MTE is equal to

$$\beta^{MTE}(p) = \frac{\partial E(Y|P(Z) = p)}{\partial p} \tag{45}$$

Thus we can estimate a nonparametric regression of Y on $P(Z)$, where $P(Z)$ is the propensity score or the probability that treatment is assigned, given Z. The estimate of the marginal treatment effect then is the derivative of this nonparametric estimator.

As a tool the MTE can be very useful. For example suppose we estimate the MTE for going to college as a function of different levels of p; and suppose that we find that the MTE is high for those with a low probability of attending college. This indicates that a policy that targets those with a low probability of attendance is likely to have high returns. In some circumstances, such an empirical finding may be interpreted as reflecting the presence of liquidity or other constraints of attending college. The difficulty with the MTE is the extent that the instruments Z can span a continuous support of the probability of being assigned to treatment between 0 and 1. It is generally difficult to identify instruments that satisfy the independence assumptions and have sufficient support. In practice many discrete instruments may serve that purpose.

3 The returns to education and labour force participation

3.1 Bias to the estimated returns when Participation is ignored

We have emphasised the issue of endogenous education when estimating the returns. However, another equally important problem when estimating the returns is that of missing wages for nonworking individuals. Comparing the wages of workers can lead to biased results on the returns to education because those with missing wages are

not selected randomly: suppose we measure the wage returns to college by

$$\Delta^{C/H} = E(\ln w|C, P = 1) - E(\ln w|H, P = 1)$$

If education affects participation in the labour market ($P = 1$) then the ability composition of those working with a college degree will be different from the ability composition for those working with a high school degree even if education is exogenous for wages. This problem is important because the proportion of nonworkers can be very high. In the UK for example in 2000 only 78% of men with statutory schooling and 85% high school graduates worked. For women the respective figures are 60% and 75%.

A simple analysis based on Heckman's (1979) model will illustrate that ignoring non-participation is likely to lead to an underestimate of the returns to education. The intuition is simple: if education increases participation the composition of the workers with higher level of education will be worse than the composition of the workers with lower levels of education, where participation is lower. This intuitive analysis is based on many strong assumptions. For example is individuals sort into educational groups by comparative advantage it is no longer obvious how the bias will go. One way of understanding the potential amount of bias is to use Manski's (1994) worst case bounds.

Denote by $F(w|ed)$ the distribution of wages for the entire population with education level *ed*, irrespective of work status. Assume for simplicity that education itself is exogenous. The concept here is that each individual has some wage they would earn were they to work; however this wage is not observed when the individual is not employed. Hence $F(w|ed)$ is not itself observed. Instead we observe $F(w|ed, P = 1)$, where $P = 1$ denotes those working. The two are related by

$$F(w|ed) = F(w|ed, P = 1)\Pr(P = 1|ed) + F(w|ed, P = 0)\Pr(P = 0|ed)$$

where $F(w|ed, P=0)$ is not observed. This implies that

$$F(w|ed,P=1)\Pr(P=1|ed) \leq F(w|ed) \leq F(w|ed,P=1)+\Pr(P=0|ed)$$

which implies that the width of the bounds for the unconditional distribution of wages is $\Pr(P=0|ed)$, giving a potentially very large range for the order statistics (median, quartiles etc.) of the wage distribution and leaving all the order statistics below $\Pr(P=0|ed)$ and above $\Pr(P=1|ed)$ unidentified. Moreover without restrictions on the support of wages it is not possible to bound the mean of wages corresponding to $F(w|ed)$, without other information or restrictions. This means that the wage returns to education are unidentified and those based on comparing order statistics, such as the medians can lie in very wide ranges. Although this is definitely not a novel point it is often overlooked when estimating returns to education.

3.2 Accounting for non-participation

The original way of dealing with the issue was that of correcting wages for selection into employment as in Gronau (1974) and Heckman (1974, 1979) and many others that followed. Heckman and Honoré (1990) provided an in-depth analysis of identification in a Roy model that includes the simple selection model. Heckman (1990), Ahn and Powell (1993), and Das, Newey, and Vella (2003) developed further identification results. The literature on selectivity corrected wage equations is vast and we will not discuss it further here.

Alternatively we discuss recent developments on using bounds to account for selection when estimating the distribution of wages or more specifically the returns to education. In an earlier section we illustrated the use of bounds for allowing for the endogeneity of wages. Here we discuss the approach of Blundell, Gosling, Ichimura and Meghir (2007) (henceforth BGIM) who develop bounds that allow us to control

for the effects of selection into work.

Worst case bounds are generally too wide to be useful, other than to illustrate that without any assumptions it is very difficult to say anything. Thus BGIM obtain tighter bounds by using three different restrictions. In the first they assume that the distribution of wages of workers either stochastically dominates the unobserved one of non-workers or at least has a higher median than that of non-workers. The idea here is that there is positive selection into the labour market. Under the stronger stochastic dominance restriction the lower bound to the distribution of wages increases and we obtain that

$$F(w|ed,P=1) \leq F(w|ed) \leq F(w|ed,P=1)\Pr(P=1|ed)+\Pr(P=0|ed).$$

This restriction is never testable. In addition it does not follow from economic theory, particularly if we do not condition on wealth. Indeed it is possible that higher wage individuals have higher reservation wages because they are on average wealthier - this would lead to a violation of the stochastic dominance assumption. On the other hand there is circumstantial evidence, presented in BGIM that there is positive selection into the labour market, even when we do not condition on wealth.

An alternative and non-nested set of restrictions relates to the use of instruments. The idea is the same as the one presented earlier for the case of endogenous education. Denote the instrument by Z. Note that while the observed distribution of wages for workers $F(w|Z,ed,P=1)$ will depend on Z in general, the population distribution of

wages $F(w|ed)$ will not, by assumption. Then BGIM show the bounds to be[30]

$$\max_z \{F(w|Z=z,ed,P=1)\Pr(P=1|Z=z,ed)\}$$

$$\leq F(w|ed) \leq \qquad . \quad (46)$$

$$\min_z \{F(w|Z=z,ed,P=1)\Pr(P=1|Z=z,ed)+\Pr(P=0|Z=z,ed)\}.$$

In the above expression we search over different values of z to identify the tightest bounds. These may not necessarily be where $\Pr(P=1|Z=z,ed)$ is maximised, unless it goes to one. In contrast to the case with stochastic dominance this restriction has some testable implications because its violation can lead to the bounds crossing, i.e. to the lower bound being higher than the upper bound; this provides a means of testing for the validity of the exclusion restriction, although note that it may be possible for the restriction to be false and still the bounds may not cross. Hence the test may not have power one against the null that wages are independent of the instrument (see the section on bounds for endogenous education).

Manski and Pepper (2000) originally presented the idea of monotone instrumental variables. In this case it is no longer assumed that wages are independent of the instrument but that the mean of wages is monotonic in the instrument. BGIM extend this idea to the entire distribution by assuming that

$$F(w|Z=z',ed) \leq F(w|Z=z,ed) \; \forall \; w, \; z, \; z' \text{ with } z<z'$$

[30] See Blundell, Gosling, Ichimura and Meghir (2007), equation 8.

This then implies the following bounds *conditional* on some value of the instrument z_1

$$F(w|Z=z_1,ed) \geq F^l(w|Z=z_1,ed) \equiv$$

$$\max_{z>z_1}\{F(w|ed,Z=z,P=1)\Pr(P=1|Z=z,ed)\}$$

$$F(w|Z=z_1,ed) \leq F^u(w|Z=z_1,ed) \equiv$$

$$\max_{z>z_1}\{F(w|ed,Z=z,P=1)\Pr(P=1|Z=z,ed)+\Pr(P=0|Z=z,ed)\}$$

By averaging over all possible values of z we can then obtain bounds to $F(w|ed)$ that are consistent with the monotonicity assumption. These take the form

$$E_z\left[F^l(w|z,ed)\right] \leq F(w|ed) \leq E_z[F^u(w|z,ed)]$$

where E_z denotes the expectation with respect to the distribution of z. As in the case of the exclusion restrictions, these bounds can cross if the monotonicity restriction is not valid. Moreover, it is possible to combine the monotonicity assumption with positive selection, such as stochastic dominance.

This procedure bounds the distribution of wages. This allows us to bound some quantiles as discussed above, but not means without support conditions on the distribution of wages. Bounding differences in the order statistics can give us a measure of the wage returns to education; this is done as in equation 41. As an indication of the results that one can obtain Table 1 presents results on bounds to the returns to college versus high school estimated by BGIM for men in the UK. These use both the monotonicity restriction and the assumption that the median wage of workers is higher of that which non-workers would earn. The instrument used is the income that an individual would have if they did not work; this is determined by the benefit system in place at the time and the demographic structure of the household. Interestingly BGIM report

Birth Cohort	Age	Bounds
1965	30	23%-29%
1955	30	18%-23%
1945	35	17%-20%
1935*	45	16%
1925*	55	44%
*Point Estimate obtained		
Source BGIM, Figure 13;		

Table 1: Bounds to the returns to college relative to high school by cohort - Males, UK

that the instrument they use is rejected when used as an exclusion restriction based on the test for crossing bounds. This may reflect the fact that characteristics determining out of work income maybe related to wages. The weaker monotonicity restriction allows these factors to be correlated with wages, so long as the distribution of wages is monotonically related to out of work income. Indeed, the monotonicity restriction is not rejected. The reported returns correspond to different ages, because older cohorts are not observed at younger ages. These bounds are quite tight and in some cases lead effectively to point estimates. This illustrates the point that identification may be obtained without having to assume that participation rates are 1 for a set of values of the instrument Z.

3.3 Non participation and endogeneity

The discussion in this chapter has dealt with the implications of endogeneity of education when estimating returns and separately with the implications of non-participation. Both these issues are very important and ignoring them can cause bias. It is thus important to deal with both issues simultaneously, although this is not always done in the empirical literature.

From a parametric view point the obvious way to proceed would be to extend the model presented earlier for educational choice to one that also allowed for labour force

participation. Two examples of such models are Keane and Wolpin (1997) and Adda, Dustmann, Meghir and Robin (2009) among others. In these models individuals decide on their educational attainment; subsequently during their labour market career individuals also decide (among other choices) whether to work or not. Within this context the models can be enriched further by allowing for endogenous accumulation of human capital in work (experience) as in Eckstein and Wolpin (1989) and the papers mentioned above among others, and by allowing for search frictions as in Adda et al. (2009) as well as earlier papers including for example Wolpin (1992). In these integrated models the issue of endogeneity of education and labour force participation is treated in a comprehensive, albeit fully parametric way. Non-parametric approaches, either to point estimation or to just bounds have not been implemented in practice to our knowledge and neither has a comprehensive analysis of identification taken place. However, preliminary unpublished calculations by Hide Ichimura and Costas Meghir suggested that bounds would be too wide to be informative, even if we were to assume that only those expecting to gain from education actually attended. Finding suitable restrictions that would make such an approach informative would be an important advance.

To provide a brief illustration of how the model of section 2.2 can be generalised to allow for non participation we rewrite the value functions for the period of working

life in period (age) t as[31]

$$V_t^{WJ}(X_{it},\tau_i^J) = w_{it}^J + \beta Emax\left[V_{t+1}^{WJ}(X_{it+1},\tau_i^J), V_{t+1}^{0J}(X_{it+1},\tau_i^J)\right] \quad \text{Value when working, } J = S,H,C$$

$$V_t^{0J}(X_{it},\tau_i^J) = b(X_{it}) + \xi_{it} + \beta Emax\left[V_{t+1}^{WJ}(X_{it},\tau_i^J), V_{t+1}^{0J}(X_{it},\tau_i^J)\right] \quad \text{Value when not working, } J = S,H,C$$

$$\log w_{it}^J = m^J(X_{it}) + \tau_i^J + \varepsilon_{it}^J \quad \text{Wages for education level } J = S,H,C$$

$$X_{it+1} = X_{it} + P_{it} \quad \text{Experience} \tag{47}$$

In 47 X_{it} represents experience, which here is defined as the number of periods working in the labour market. We have introduced a flow value for leisure, $b(X_{it})+\xi_{it}$ which is stochastic and depends on experience, possibly reflecting the level of benefits dependent on past wages, or contributions to some unemployment insurance fund. Wages are assumed to depend on work experience and education J but on on age. The decision to work is

$$P_{it} = 1(V_{it}^{WJ} > V_{it}^{0J})$$

Experience X_{it} is endogenous, because it is an accumulation of past work decisions, which depend on the unobserved heterogeneity component τ_i^J. The terminal value function could be specified as in 10, where X now represents experience at the time when we stop having information on wages. To complete the model one needs to specify the stochastic properties of the shocks ε and ξ including their distribution. Because of the finite life nature of the model the value functions depend on age t as well as on experience. Thus the model is solved backwards from the terminal point to the point where the education decisions are made. The work value functions (V^{WJ},$J = S,H,C$) are evaluated at zero experience when solving for the education choice. In this model it is assumed that all education decisions are made at the beginning of the

[31] Recall that the education levels are S for statutory, H for high school and C for college.

lifecycle, but there is no reason why we cannot further generalise the model to include the possibility of returning to full time education after a period of work. Finally, note that the return to education, as seen at the beginning of the lifecycle, will be generalised now to include the effect of education on the length of work spells.

Estimation will be similar to that discussed in section 2.3 with the important modification that we need to model the probability of working in each period and we need to account for the fact that wages are observed for workers only. Assuming for simplicity that $\varepsilon$and ξ are iid normal the likelihood contribution for an observed career of T_i periods, conditional on education and on unobserved heterogeneity becomes

$$L_i^J \equiv L(w_{is}...w_{iT_i}|J_i,\tau_i,X_i,\theta)$$

$$= \prod_{t=1}^{T_i}\left[g^J(w_{it}|\ X_i,\ \tau_i^{J_i})Pr(P_{it}=1|\ X_i,\ \tau_i^{J_i},w_{it})\right]^{P_{it}=1}\times \tag{48}$$

$$\left[\int_w g^J(w|\ X_i,\ \tau_i^{J_i})Pr(P_{it}=0|\ X_i,\ \tau_i^{J_i},w)dw\right]^{P_{it}=0}$$

The construction of the rest of the likelihood follows as in section likelihood follows as in section 2.3.In particular 16completes the likelihood function including the step relating to educational choice. The unobserved heterogeneity is integrated out and the likelihood for the whole sample is put together as in 17 The joint distribution of $G(F,\tau)$ accounts for the endogeneity of education in wages and participation over the subsequent periods of the lifecycle.

Finally, we have already discussed the difficulties relating to nonparametric identification of such models, even without endogenous participation. Obviously with participation being endogenous matters do not become easier because we would now need to also identify the distribution of the shocks to leisure ξ as well as the distribution of the shocks to wages. In practice many of these aspects will be specified parametrically. However, identification is aided by the presence of exogenous variation at the time of

education choice. For example continuous (or even discrete) variables that affect the costs of education and vary exogenously across individuals, can provide credible exogenous sources of identification. Attanasio, Meghir and Santiago (2009) argue that the use of a randomized experiment with say educational incentives, as in the PROGRESA conditional cash transfer programme in Mexico can serve such a purpose.

4 Education Policy and the estimated returns to Education

We have argued that it is hard to interpret results from estimating wage equations without a theoretical foundation. The need for models is reinforced when we consider scaling up human capital policies. Consider for example the impact of a school subsidy for children from low income families.[32] Estimating the impact of such a policy on a small scale is insufficient for understanding its longer term effects, even if we can estimate the wage returns to education for those induced into education due to the policy. Among other issues, we need to know the mechanism through which the subsidy acted: was the increase primarily due to a distortion of incentives or due to the alleviation of liquidity constraints? Second we need to know what the general equilibrium effects are. The latter include i. the effect of changes in the supply of skill on skill prices, which

[32] An example of such a policy is the Education Maintenance allowance in the UK, a subsidy to 16 year olds for post-compulsory school attendance. in the UK, evaluated by Dearden, Emmerson, Frayne and Meghir (2009). This is a conditional cash transfer offered to pupils who completed statutory education at 16 and whose family's income is low, on condition they remain in full time education At the time the policy was evaluated in 1999-2001, the amount received was at a maximum when the family earned less than $20,800 a year and thereafter declined linearly up until the family income reached $48000, which was the eligibility threshold. Based on a pilot/control comparison using matching the estimated effect of the policy was to increase post-compulsory school participation by 6-7 percentage points for eligible children.

can feed back on to the decision to obtain the extra education; ii. the potential dilution of education quality as the resources are spread out more thinly; and iii. the peer effects of keeping more of the 16 year olds in school, which could change the composition of the classrooms as well as the cultural norms. Allowing for all this is a tall order and it will be hard, to say the least, to build a credible model that will be able to capture all these elements. One needs to make some realistic choices of which of these aspects are likely to be of first order importance. We will briefly discuss dealing with the changes in human capital prices as a result of the increased supply of educated workers due to the subsidy. Important work in this field of general equilibrium models with heterogeneous agents has been carried out by Heckman, Lochner and Taber (1998), Lee (2005), Lee and Wolpin (2006), and Gallipoli, Meghir and Violante (2008).[33]

Start with the model in section 2.2 and assume this has been estimated on data either collected as part of the experimental pilot or from observational survey data. This is the first building block. Since the problem has been set up as a lifecycle one, to solve for equilibrium one needs to set it up an an overlapping generations model. However, since the environment is stationary the problem to be solved for each generation is identical. Thus for any given set of human capital prices we can solve the individual problem and then account for the number of individuals in each education group at each point in time.

The next step is to estimate the production function of the form presented in section 2.1. Heckman, Lochner and Taber (1998) for example estimate a production function with two human capital inputs: less than college and college, while Gallipoli et al (2008) allow for three human capital inputs as in the model of sections 2.1 and 2.2. Both allow for one factor of capital. The estimated production function and the assumption of competitive labour markets allows us to derive relative human capital prices for each group as as function of employed human capital. Both authors estimate

[33] In Gallipoli et al (2008) we also consider other important issue, such as the role of parental transfers liquidity constraints.

substitution elasticities for the human capital inputs that imply quite a lot of sensitivity of relative prices to changes in supplies.[34]

Simple policy simulation would then compare the baseline outcomes (essentially the data) to the results from a simulated new steady state arising as a result of implementing a new policy such as an educational subsidy. The simulation based on the model described here would allow for the effect of changes in individual incentives; for the impact of funding the subsidy by raising taxation through say income taxes, which would compress the effects of education on wages; and for the effect of changes in the return to education induced by a new equilibrium in the labour market as supplies change.

The point is that all these effects can be potentially important and the results can be sensitive to the assumed environment and the specification of the model. Heckman, Lochner and Taber (1998) show that the GE effects can almost neutralise the effects of a policy. Lee (2005) has a different model specification where the feedback effects from GE are small. This shows that the results can be sensitive to important modelling choices and that we do need to know all the components of the model to acquire a good understanding of what the policy will achieve. Just estimating the wage returns to education, even when we can do so based on a partial model, is useful but is only part of the story as far as design of policy is concerned.

[34] This may be controversial because in an open economy with more goods than factors, trade can lead to factor price equalisation. If this was really the case then policies that changed the supply of human capital would not affect human capital prices and would not be subject to such general equilibrium effects. Nevertheless, factor price equalisation is either a very slow process or is prevented by other mechanisms.

5 Estimation of School Input Effects

Variation in the return to observed schooling comes from many sources, one of the most important of which is the quality of education. A growing body of research investigates the effects of various educational inputs, and the proliferation of administrative and survey data facilitates such analyses. Similar to research on school attainment, endogenous choices and unobserved heterogeneity complicate efforts to identify variable effects. Empirical models must explicitly or implicitly account for the inter-related choices of families, teachers, administrators and policy makers to avoid contamination from confounding factors. In this section we discuss selected papers covering four topics that have generated substantial interest in education research in order to highlight key empirical and methodological issues including the treatment of the multiple decisions that determine the allocation of educational inputs. These areas are class size effects, teacher quality, housing market capitalisation of school quality, and the effects of choice and accountability. The selected papers use a variety of approaches and types of data, and we emphasise implications of the specification choices. Methods discussed include the use of observed characteristics as controls, various types of IV techniques including regression discontinuity and lottery generated quasi-random assignment, difference-in-differences and large-scale fixed effects specifications, hierarchical linear modelling (HLM), and structural discreet choice models. As in the previous section, we highlight the explicit and implicit assumptions regarding the underlying choice framework as they relate to both the identification and interpretation of variable effects. Prior to considering these four topics we describe a model of housing choice that highlights many of the determinants of family location decisions and then discuss some general issues related to learning and the accumulation of human capital. The latter discussion focuses on empirically relevant issues pertinent to much education research.

5.1 Housing Choice[35]

Consider the location equilibrium of a household that resides at location d*. Ignoring mobility costs, the household will be in equilibrium at location d^* if:

$$d^* = argmax_d E \int_H \left[U(X^d_\tau, SQ^d_\tau, O^d_\tau | w^d, f, \kappa_d, p_d) d\tau \right] \; for d \in \{d\}$$

where expected utility is accumulated over the relevant planning horizon, H, and the location, $d*$, is chosen once and for all (for simplicity) compared to all d.[36] Each location is associated with a wage w_d, preferences for such a location k_d, prices p_d (which include house prices), a set of local amenities O and in particular SQ is school quality. Utility may depend on individual abilities f, that drive wages. They also depend on a vector of household consumption, on labour supply and on demographics, all of which have all been maximized given location.[37] Hence $U(.)$ represents indirect utility given location. Largely static variants of this lie behind general theories of urban location decisions, the quality of local public services, and the demand for local government services (cf. Straszheim, 1987; Tiebout, 1956; Wildasin, 1987). In the simplest models a household optimizes equation 1 across all of the feasible locations within its choice set given complete information for all periods. Yet lifecycle changes, unexpected shocks or incorrect predictions move families out of equilibrium and often lead to relocation. For example, households may decide to relocate because of changes in expected lifetime income, family structure (additions of children, divorce or remarriage), perceptions of the quality of local public services including schools, the distribution of employment opportunities, or other factors. Even in the absence of prediction error, rising income may reduce borrowing constraints and expand opportunities. Note further that moving costs introduce inertia into the decisions, so that at any point in time a household might drift away from its current utility maximising

[35] This section draws from Hanushek, Kain and Rivkin (2004)
[36] See Kennan and Walker (2009)
[37] We are being vague about household formation so as not to complicate the notation.

location and might not move until a time when the utility loss from $d*$ compared to the next best alternative becomes large. A much more complete model of location choice is developed by Kennan and Walker (2009) and shows the complexity of such decision making. Developing this in a general equilibrium framework and understanding the interaction between amenities, labour market opportunities and preferences is at the heart of understanding how individuals end up choosing school quality as a function of preferences and abilities as well as prices. This simple model highlights some of the main impediments to the estimation of school and teacher effects. First, it is quite difficult to account for all the factors that lead families to make different location and school choices. Second, even with panel data to account for student or family fixed effects, changes sizeable enough to induce geographic moves are likely also to have direct effects on outcomes. This raises immediate questions about the validity of exogeneity restrictions in panel data analyses.

5.2 Learning Dynamics

The cumulative nature of knowledge acquisition introduces an additional complication into the estimation of school and teacher effects, and we now consider specification issues directly related to the modelling of the dynamics of knowledge retention in a data generating process driven by the multiple dimensions of choices that determine the distribution of teacher and school characteristics. In order to highlight the specification issues related to assumptions regarding the rate of knowledge depreciation we assume no heterogeneity in school input effects. Equation 49 models the outcome of student i in year t as a function of a school input S, a vector of control variables X, a time varying student effect α_{it} that captures unobserved student heterogeneity other than differences resulting from S and X, and an error term e that represents all other determinants of A including measurement error in the outcome variable and unobserved

school, community, and family influences.

$$A_{it} = \alpha_{it} + \beta X_{it} + \delta S_{it} + e_{it} \tag{49}$$

The variables X and S have t subscripts, because many studies make use of panel data that contain multiple measures of family and school variables. In other cases including research on earnings, earnings in year t are regressed on family and school characteristics measured during childhood.

If S is orthogonal to α and e estimation of Equation 49 produces consistent estimates of δ. In reality, OLS estimation of Equation 49 is unlikely to produce consistent estimates given limited information available and the complex processes that determine the distribution of school and teacher characteristics. These include 1) family location and schooling decisions that are part of the previously discussed process of life-cycle optimisation of utility; 2) utility maximising choices of schools and districts by teachers and other school personnel; 3) purposeful matching of students and teachers in classrooms; and 4) the political and judicial processes that determine school finances and a range of laws that affect the allocation of resources and students among classrooms, schools, and districts.

Although confounding factors and consequent omitted variables bias tend to be the primary issue considered in the discussion of the merits of most empirical approaches, other measurement and specification errors also threaten estimation of equation 49. As we discuss below, complications introduced by test-measurement error have received considerable attention in studies of teacher and school effects and related policies including merit pay and NCLB accountability.[38]

Another frequently discussed specification issue is the appropriate treatment of the history of family and school inputs given life cycle utility optimisation. As Cunha, Heckman, and Schennach (2010), Cunha and Heckman (2008), and Todd and Wolpin

[38]The arbitrary normalisation of test scores also complicates estimation, as monotonic transformations may lead to very different findings. Cunha and Heckman (2008) discuss this issue.

(2003) emphasise, the development of cognitive and non-cognitive skills is a cumulative and complex process and the failure to account for the history of inputs can lead to biased estimates of the effects of variables of interest. Limited availability of historical data on parental, school, and community inputs and the endogeneity of parental inputs impede efforts to estimate the full life-cycle model, leading to the use of lagged achievement measures and student fixed effects to account for the history of parent, community, and school input effects. Such methods fail to capture the nuances of skill development processes involving endogenous parental behaviour in which there appear to be sensitive periods during childhood for parental investments in both types of skills (Cunha and Heckman, 2008).

Nonetheless, the value-added and fixed effect methods may account for the aggregate effects of the history of inputs, and it is informative to examine the implications of imposing various assumptions on the rate of knowledge depreciation. Therefore we assume no confounding family or other factors including endogenous responses to realised school quality and describe the implications of five commonly used approaches using a simple model of achievement for student i in grade G in which a constant proportion (1-θ) of knowledge is lost each year ($0 \leq \theta \leq 1$), the effect and variance of SC do not differ across grades, the covariance of SC across grades is constant, and the error is orthogonal to SC in all periods.

$$A_{iG} = \beta \sum_{g=0}^{G} \theta^{G-g} SC_g + error \tag{50}$$

First consider an OLS regression of achievement in grade G on SC in grade G with no control for prior achievement and thus likely to be subject to the influences of confounding factors. In this case the error includes effects of all past values of the school characteristic:

$$A_{iG} = SC_G \beta_{level} + \beta \sum_{g=0}^{G-1} \theta^{G-g} SC_g + error \tag{51}$$

and[39]

$$E(\hat{\beta}_{level}) = \beta_{level} + \beta\rho\left(\sum_{g=0}^{G-1}\theta^{G-g}\right)\frac{1}{var(SC)} \tag{52}$$

where ρ is the covariance of the school characteristic in grades i and j that is assumed not to vary by number of years or grades apart.

In general, the magnitude of any bias depends on both θ and ρ. Not surprisingly, bias decreases along with the rate of decay and approaches zero as θ approaches zero. In the special case of random assignment experiments, IV, or other methods that isolate the component of the school characteristic in grade G that is uncorrelated with the school characteristic in other grades, this specification produces unbiased estimates of β regardless of θ.

Any correlation between the current and past values of the school characteristic complicates interpretation of the estimate and limits the generalisability of the findings. The estimate would reflect some weighted average of current and depreciated past effects, where the weighting depends upon the often unknown serial correlation in the school characteristic. Given the substantial demographic differences in school mobility rates by race, ethnicity, and family income, estimates would tend to be higher for students in stable schools even if the true effects were either similar for all students or higher for those in more turbulent environments.[40]

More compelling approaches use multiple years of test score results to account for student heterogeneity. One such approach is the student fixed effects model without a control for lagged achievement. Taking first differences of equation 49 (subtracting A_{G-1} from A_G) to remove any student fixed error component gives:

[39]Given the assumption that θ is a constant, the expected value can be calculated using the omitted variables bias formula treating the terms in the summation as a single variable.

[40]Hanushek, Kain, and Rivkin (2004) describe mobility differences in Texas elementary schools.

$$A_{iG} - A_{iG-1} = (SC_G - SC_{G-1})\beta_{f.e.} + \beta\left[\theta SC_{G-1} + \sum_{g=0}^{G-2}(\theta^{G-g} - \theta^{G-(g+1)})SC_g\right] + error \quad (53)$$

In this case[41]

$$E(\hat{\beta}_{f.e.}) = \beta - \beta\theta/2 \quad (54)$$

In contrast to the simple levels model, the fixed effect specification produces an estimate of the school characteristic effect that is biased toward zero as long as the rate of decay (1-θ) does not equal one. Notice that the magnitude of the bias does not depend upon the value of correlation of the school characteristic across grades, ρ.

A second approach uses prior year test score to account for student heterogeneity by subtracting it from current year score and using the test score gain as the dependent variable. This model is a special case of the value added model of test score in grade g regressed on test score in grade g-1 and the school characteristic in which θ is assumed to equal 1. This model is often preferred to the unrestricted value added model, because the inclusion of an imprecisely measured lagged endogenous variable as a regressor can introduce other types of specification error including errors in variables and endogeneity bias.[42] Here achievement in grade g minus achievement in grade g-1 is regressed on the school characteristic in grade G.

$$A_{iG} - A_{iG-1} = SC_G\beta_{gain} + \beta\sum_{g=1}^{G}(\theta^g - \theta^{g-1})SC_{G-g} + error \quad (55)$$

[41]Here the term in brackets is treated as the single omitted variable. Notice that the assumption of a constant covariance regardless of the number of grades between grades i and j mean that the covariance between (SC_G–SC_{G-1}) and SC_g equals zero for all values of g less than G-1.

[42]Numerous studies use test score gain as the dependent variable including Rivkin, Hanushek, and Kain (2005) and Harris and Sass (2009); both papers discuss the model in some detail.

and[43]

$$E(\hat{\beta}_{gain}) = \beta - \beta\frac{(1-\theta^G)\rho}{var(SC)} \tag{56}$$

The magnitude of any bias depends on both θ and ρ. In the case of θ, the problem is that the violation of the assumption of no knowledge depreciation means that the higher the lagged score the higher is the over-estimate of expected test score in the current year. Not surprisingly given the structure of the model, bias decreases as the true value of θ increases and disappears if there is no loss of knowledge from year to year, i.e. θ=1. As is the case with the levels model, the use of a value added framework, random assignment experiments, IV, or other methods that isolate the component of the school characteristic in grade G that is uncorrelated with the value of the characteristic in other grades produce unbiased estimates of β regardless of the true value of θ; If ρ equals zero the error introduced by the mis-specification is orthogonal to the school characteristic and does not introduce bias.

As is the case with the model with lagged achievement, the inclusion of a student fixed effects does not eliminate the specification error to the gains model. Taking first differences of equation 55 (subtracting $A_{G-2} - A_{G-1}$ from $A_G - A_{G-1}$) to remove any student fixed effect in gains gives the fixed effect in gains model.

$$(A_{iG} - A_{iG-1}) - (A_{iG-1} - A_{iG-2}) =$$

$$(SC_G - SC_{G-1})\beta_{f.e.gain} + \beta\left[(\theta-1)SC_{G-1} + \sum_{g=2}^{G}(\theta^g - 2\theta^{g-1} + \theta^{g-2})SC_{G-g}\right] + error \tag{57}$$

and[44]

[43]Given the assumption that $cov(SC_i, SC_j)$is constant regardless of the number of grades apart, all terms cancel except for one grade $G-1$ term and one grade 0 term.

[44]Given the assumption that $cov(SC_i, SC_j)$ is constant regardless of the number of grades apart, all terms

$$E(\hat{\beta}_{f.e.gain}) = \beta + \beta(1-\theta)/2 \tag{58}$$

Similar to the case for levels, bias in the fixed effect in gains specification is of the opposite sign as any bias in the gains model without student fixed effects, and the magnitude of the bias does not depend upon the value of ρ. Notice that the bias is the same magnitude but the opposite sign for the two fixed effects models in cases where θ=0.5.

In summary, this simple education production function model illustrates that the violation of a strong assumption regarding the rate of knowledge depreciation introduces bias in fixed effects models regardless of the magnitude of ρ and in models without fixed effects as long as the magnitude of ρ does not equal zero. This provides some rationale for the use of models including those with student fixed effects in which the value of θ is not constrained to equal zero or one.[45]

Of course the dynamics of non-cognitive and cognitive skill formation necessitate the use of richer empirical models to control for potentially confounding family and community influences. Cunha and Heckman (2008) argue in favour of the use of a latent variable framework to account for the endogeneity of parental inputs and multiplicity of potential proxies for family background. Although their interest in the pattern of family effects differs from our focus on school inputs, the issues of endogenous parental behaviour and student heterogeneity along multiple dimensions have direct relevance to the identification of school input effects given the possibility of parental responses to realised teacher quality, class size, and other school inputs and potential non-random sorting of students into schools and classrooms on the basis of cognitive and non-cognitive skills.

Another important issue is the scale of the test scores. This is arbitrary and any

cancel except for one grade G-1 term and one grade 0 term.

[45]Nerlove (1971), Nickell (1981), Arellano and Bond (1991) and Hsiao (2003) discusses estimation fixed effects models with a lagged dependent variable and biases that can arise from within groups estimation.

monotonic transformation provides the same information. However, this raises two related issues: first there is no reason to expect that linearity and additivity of fixed effects or other unobservables, such as the one postulated in equations (49, 50, 51 or 55) has to be valid for the particular scale we happen to be working with. Second, the comparison of changes in test scores is not invariant to monotonic transformations: for example, the statement that the change in test scores was larger for some group than for another is not invariant to monotonic transformations of the scores. One way around this problem is to find a natural cardinalisation or anchoring of the test scores.[46] This in itself raises interesting questions, because test scores are an aggregation of answers to many different questions and also provide the benchmark for the teacher to evaluate her success. As a result the way test scores are arrived at may affect teacher incentives. Thus suppose for example that test scores were calculated in order to maximise their predictive power with respect to future wages, academic attainment or other outcome. Then the objective for teachers could be defined as aligned with such longer term outcomes, and the scores could be anchored in that metric. This would not eliminate complications introduced by non-linearities, but at least it would fix the metric and define clearly the meaning of the linearity assumption. The difficulty is of course that we often do not have a clearly measured link between the test scores at hand and an outcome variable let alone agreement on what that outcome variable should be.

5.3 Estimation of Class Size Effects

Similar to the case of the return to schooling, the benefits of smaller classes may vary along several dimensions including initial class size, student characteristics, the school environment, and the nature of the comparison. Yet little of the empirical work on class size is grounded in a conceptual model of class size effects that points toward

[46] see Cunha, Heckman and Schennach (2010)

particularly types of heterogeneity. Rather analyses typically provide average effects or effects that differ by demographic group or grade. If there is substantial heterogeneity in the benefits of smaller classes, samples drawn from different populations would be expected to produce different estimates of average effects for all students or even students in a particular demographic group. Estimates would also be expected to differ on the basis of whether or not the differences in class size used to identify a coefficient are related to differences in unobserved teacher quality resulting from either teacher preferences for smaller classes or any expansion in the number of teaching positions necessary to reduce class size.

We begin this section by outlining a model of the relationship between learning and class size and then discuss six studies of class size effects that use a range of methods and types of data. The model allows for heterogeneity by student ability and the level of disruptive behaviour in a classroom and can incorporate general equilibrium effects resulting from changes in teacher quality. Recent research finds evidence of heterogeneous class size effects along the achievement distribution, and Lazear (2001) highlights differences in the level of disruption as a likely explanation for why lower income students appear to realise larger benefits from smaller classes.[47]

5.3.1 Model[48]

Equation 59 models learning for student i in classroom c in school j as a function of the amount of classroom time available for learning and the value of that time in terms of the quality of the teaching and relevance of the material, plus all other student, community, and school factors:

$$learning_{icj} = \rho(d)^{n}_{cj} q(n,a)_{icj} + X_{icj} \tag{59}$$

[47] Ding and Lehrer (2005), Konstantopoulos (2008), and McKee, Rivkin and Sims (2010) all find that the benefits of smaller classes appear to increase with achievement.

[48] The discussion is drawn from McKee, Rivkin and Sims (2010).

where ρ is the proportion of time a student is not disrupting the class; d is the classroom average propensity to disrupt the class; q is the value of a unit of instructional time; n is class size; a is an index of ability[49] and X is a vector of other student, community and school factors.

The term $\rho(d)^n_{cj}$ is drawn from Lazear (2001) and represents the share of class time not lost to disruption by any of the n students in the room, while the term $q(n,a)_{icj}$ models the value of a unit of instructional time as function of both class size and academic preparation. Variation in classroom behaviour, d, and academic preparation, a, provide two dimensions of potential heterogeneity in the benefit of smaller classes. Importantly, all students in a classroom experience the same amount of instructional time, but the value of instructional time may vary by ability due to targeting of the curriculum, the distribution of teacher effort, and student heterogeneity.

In order to illustrate the ways in which disruption and academic preparation may affect the benefits from class size reduction, we take the derivative of equation 1 with respect to n and then again with respect to d (equation 60) and a (equation 61 below):

$$\frac{\partial^2 learning}{\partial n \partial d} = \left\{ [\rho(d)]^{n-1} \frac{\partial \rho(d)}{\partial d} \right\} \left\{ (n \ln(\rho(d)) + 1) q(n,a) + n \frac{\partial q(n,a)}{\partial n} \right\} \quad (60)$$

Equation 60 illustrates the relationship between the propensity to disrupt class and the benefit of class size reduction.[50] The product of the two relationships in squiggly brackets determines the sign of the cross-partial derivative of learning with respect to n and d. The first is negative, as the derivative of ρ with respect to d is assumed to be negative (a higher average propensity to disrupt reduces the share of time available for learning), while the second is ambiguous and depends on the magnitudes of the various

[49] Ability represents a one dimensional index of academic skill and is not meant to refer to capture innate differences.

[50] The derivative of learning with respect to class size (n) equals $[\rho(d)]^n ln(\rho(d)) q(n,a) + [\rho(d)]^n \frac{\partial q(n,a)}{\partial n}$. Taking the derivative of this relationship with respect to d produces Equation 60 and with respect to a produces equation 61.

terms: $\frac{n\partial q(n,a)}{\partial n}$ is assumed to be negative: the quality of instructional time declines as class size increases for a number of reasons including more difficulty differentiating the curriculum to account for variation in academic preparation. $q(n,a)$ is positive and the product of $ln(\rho)$ and n lies between 0 and -1, so $nln(\rho)+1$ is also positive.[51]

Thus the relationship between the benefits of class size reduction and the degree of disruption (d) thus depends upon the magnitudes of two counteracting effects. First, as Lazear (2001) points out, at lower values of ρ reduced class size has a larger effect on the share of time available for learning and thus a larger effect on achievement. Second, at lower values of ρ any improvement in the quality of instruction time due to smaller classes has a lower overall impact, because classrooms with lower values of ρ have less time for learning.

Equation 61illustrates the relationship between initial achievement and the benefit of class size reduction:

$$\frac{\partial^2 learning}{\partial n \partial a} = [\rho(d)]^n \left\{ ln(\rho(d)) + \frac{\frac{\partial^2 q(n,a)}{\partial n \partial a}}{\frac{q(n,a)}{\partial a}} \right\} \tag{61}$$

As is the case with disruption, the relationship between the benefit of class size reduction and initial achievement cannot be signed a priori in this framework. Here the sum of the two relationships in squiggly brackets determines the sign of the cross-partial derivative of learning with respect to n and a. The first term, roughly the average disruption of a single student, is negative, while the ratio can be positive or negative depending upon the relationship between achievement and the quality of instruction and the relationship between achievement and the change in the quality of instruction as class size falls.

In sum, the pattern of heterogeneous effects along both dimensions cannot be predicted a priori. Moreover, differences in district policies may produce variation across

[51] At a value of ρ below 0.95, $nln(\rho)+1$ becomes negative, but at such a low value of ρ the share of class time available for instruction is well below 50 percent.

districts in the distribution of treatment effects across each of these dimensions. Finally, any accompanying changes in teacher quality may affect both the quality of instruction and level of disruption per student (changes in teacher skill at managing the classroom), and any such affects may vary by school characteristics.

5.3.2 Discussion of Empirical Analyses

Table 5.3.2 lists the six studies that we consider and describes their methods, data, and findings. The studies use a range of empirical methods to account for potential confounding factors including controls for observables, regression discontinuity, and fixed effects, and various types of administrative, survey, experimental data. Importantly, these different approaches also alter the interpretation of the parameter estimates.

Pong and Pallas (2001) use TIMSS data on the mathematics achievement of 13 year olds in nine countries to estimate the effects of class size on achievement and the degree to which curriculum and classroom instruction mediate those effects. In order to explicitly account for the multi-level structure of the data that has test score and family background measured at the individual level and class size and other school characteristics measured at the school level, the paper uses hierarchical linear modelling (henceforth HLM) estimation methods[52] The inclusion of a random school effect in the empirical model accounts for the fact that class size is a school level variable.

The results reveal little evidence of a significant negative relationship between class size and achievement in any of the countries; in fact the class size coefficients are as likely to be positive as negative. Note that the study does not account for student heterogeneity explicitly with either student fixed effects or measures of prior achievement, and it includes only a handful of family characteristics as controls. The fact that the inclusion of a small number of other school level variables tends to reduce the magnitude

[52].See Raudenbush and Bryk (2002) for a comprehensive description of HLM including two and three level random effects models.

Table 2: Selected Research on Class Size Effects

	Method	Interpretation	Data	Findings
Dearden, Ferri, and Meghir (2002)	Matching	ceteris paribus effect plus effect of differences in supply of teacher quality related to class size	UK National Child Development Study	smaller classes increase future wages of women but not men
Pong and Pallas (2001)	Hierarchical Linear Modelling	ceteris paribus effect plus effect of differences in supply of teacher quality related to class size	TIMMS	little evidence of class size effect on achievement
Pong and Pallas (2001)	State and Cohort Fixed Effects	ceteris paribus effect plus any change in state average teacher quality related to changes over time in average class size	US Census PUMS	smaller classes increase return to education in labor market
Rivkin, Hanushek and Kain (2005)	student, school by grade and school by year fixed effects	ceteris paribus effect	Texas public school administrative data	smaller classes increase achievement in 4th and 5thgrade; effects slightly larger
Angrist and Lavy (1999)	regression discontinuity	ceteris paribus effect	Israeli public school administrative data	for low income students smaller classes increase achievement in 5th grade
Krueger (1999	random assignment experiment	ceteris paribus effect	Tennessee STAR experimental data	significantly increase achievement in early grades; larger effect for lower income students

of the positive coefficients provides evidence that confounding variables introduce upward bias. Moreover, the cross-sectional estimator used in this analysis in combination with the local nature of teacher labor markets means that class size effects capture any related differences in teacher quality: teacher quality may be lower in schools with smaller classes because of the need to hire additional teachers, or teacher quality may be higher in schools with smaller classes because higher wealth communities can afford both smaller classes and higher teacher salaries.

Although the random effects provide a standard error correction to the clustering of students in schools, the validity of random effects models in general and HLM models as a special case rests in the assumption of orthogonality between the included variables and random effects. Given the limited number of covariates and multiple dimensions of choices that generate the distribution of class size, this assumption is likely to be violated in this case. Fixed effects provides an alternative to random effects that does not require the orthogonality assumption, but fixed effects do not provide a plausible approach in this case because of the absence of class size variation within schools and the assumption of linearity.[53]

Dearden, Ferri, and Meghir (2002) use rich longitudinal data that follow all subjects living in Great Britain who were born during the week of March 3-9, 1958 to estimate the effects of the pupil-teacher ratio at age 11 and at age 16 on educational attainment and wages for men and women. Equation 62 models log wage as a function of a time varying student effect α that captures unobserved differences among students, a vector X of family and community characteristics, the primary and secondary school pupil teacher ratios PPT and SPT, other included components of school and peer group quality in primary and secondary school PS and SS, and a random error.

[53] Blundell and Windmeijer (1997) show that the random and fixed effect multilevel estimators are equivalent when group sizes (in this case number of students per school) are large. However, the number of students in classrooms and schools is not large enough to eliminate the bias introduced by correlation between the random effects and included variables.

$$w_{iy} = \alpha_{iy} + \beta X_{iy} + \delta_p PPT_i + \delta_s SPT_i + \lambda_p PS_i + \lambda_s SS_i + e_{iy} \tag{62}$$

If the two pupil teacher ratio variables were uncorrelated with e and α, OLS would yield unbiased estimates of δ_p and δ_s. But as noted above, the endogeneity of family choice of school and the dependence of school finance on a number of factors including family demographics in combination with existing evidence on peer, teacher, and school effects on achievement strongly suggest that typically available variables contained in X and S will not account adequately for potentially confounding factors, thereby introducing bias into OLS estimates of δ_p and δ_s based on cross-sectional data.

However, the array of test results available in the longitudinal data along with of extensive information on schools and communities permits the inclusion of both earlier test scores as controls for unobserved heterogeneity α and a set of school and community variables to account for potentially confounding factors captured by the error.[54] In addition, the use of the pupil-teacher ratio at the school level rather than the size of individual classes circumvents potential bias introduced by the purposeful allocation of students into classes. Although a portion of the between school variation in the pupil-teacher ratio results from differences in the numbers of special education teachers and additional financing for disadvantaged populations, the included test scores and school variables should account for much of the variation in school circumstances.

It is not possible to prove that even an extensive set of controls fully accounts for all confounding factors. Altonji, Elder and Taber (2005) discuss selection on observables and unobservables and develop an informal method for assessing the probability that selection on unobservables introduces substantial bias. In this case the estimates show little sensitivity to changes in the specifications, suggesting that selection on unobservables is unlikely to introduce substantial bias.

Error in the measurement of the pupil-teacher ratio provides an additional poten-

[54] This approach is a form of matching on observable characteristics.

tial source of bias in many studies such as this where there is only a single snapshot of school characteristics to represent the school environment for a number of years. The use of a class size in a particular grade likely introduces attenuation bias, though the similar limitations in the measurement of controls also introduce bias that may amplify or offset the bias resulting from the measurement of class size. Given the relatively small size of the sample and noisiness of the wage measure, it may be difficult to identify small but educationally and economically meaningful effects such as the effects of the pupil-teacher ratio on educational attainment which are insignificant statistically but large enough to be meaningful for education policy.

In terms of interpretation, the use of pupil-teacher ratio does introduce some uncertainty, as reduction achieved through the addition of special education or intervention teachers is likely to produce a different effect on average achievement and have different implications for the educational attainment and earnings distributions than a reduction brought about by the hiring of additional classroom teachers. Therefore the estimate captures differences of the type experienced by students in the March, 1958 cohort. If a lower pupil-teacher ratio increases the supply of teacher quality, the estimated benefit of smaller classes will incorporate class size induced differences in teacher quality given the absence of information on teachers.

Concerns about omitted variables bias even in rich specifications have contributed to the expanded use of instrumental variable and fixed effect methods to account for unobserved influences. Card and Krueger (1992) provide a prominent example of an analysis with little or no information on family background that uses fixed effects for both the state of birth and state of residence to account for unobserved influences on earnings that could contaminate estimates of the pupil-teacher effect on the return to schooling. The two step procedure begins by estimating separate returns to education for each cohort-state combination from a regression of log(wage) on state of birth dummy variables, state of residence dummy variables, education by region interactions,

and separate education variables for each cohort-state of birth combination. Then the coefficients on the separate education terms are regressed on the cohort-state average pupil-teacher ratio, the cohort-state average teacher salary, and the cohort-state average school year length. Note that the use of cohort average school characteristics mitigates the measurement error introduced by a single snapshot, though any heterogeneity in class size effects by grade raises questions about the interpretation of coefficients on characteristics that aggregate information across elementary and secondary grades.

The fixed effects model uses inter-state movers and within state variation over time in school inputs to identify variable effects. Three key assumptions underlying the analysis are 1) school quality affects earnings via the return to education only; 2) selective migration does not contaminate the estimates; and 3) unobserved school or community factors are not related to the pupil-teacher ratio. Betts (1995) and Heckman, Layne-Farrar, and Todd (1996) raise questions about the first assumption, and Heckman, Layne-Farrar, and Todd (1996) document evidence of selective migration.[55] In terms of Equation 62, selective migration introduces bias by leading to a correlation between unobserved heterogeneity (represented by α in Equation 62) and the school quality measures thereby violating the exogeneity condition required for identification.

Finally, Hanushek, Rivkin, and Taylor (1996) find that the use of information aggregated to the state level affects the magnitude of omitted variables bias. In terms of Equation 62, the omission of relevant variables introduces a correlation between the school quality measures and the error which would violate the exogeneity condition required for identification. Aggregation may dampen or exacerbate any specification error depending upon the structure of the covariance between the omitted variable, the school variable of interest and the outcome. For example, if the omitted factors were to vary only at the level of aggregation (such as would be the case if these are state policy factors) and the factors were positively related to school quality and negatively related

[55] Betts (1995) investigates the effect of school resources using the NLSY and fails to find a significant relationship between wages and the pupil-teacher ratio for various parameterisation of the pupil-teacher ratio effect.

to class size, aggregation would tend to amplify the omitted variables bias. Note that the oft-asserted concern that aggregate measures of school inputs introduce measurement error, because of the difference between actual school values and the state average is incorrect regardless of whether data are aggregated to the level of aggregation of the school input measures. Rather aggregation alters the variation used to identify the estimates and does not introduce bias if the relationship between the outcome and input are linear and there are no other specification errors.[56]

This paper also uses the pupil-teacher ratio and therefore the estimates provide information on differences produced by similar underlying variation in instructional staff. However, changes in the number and potentially the quality of teachers likely accompany state average changes in the pupil-teacher ratio, conditional on salary. Therefore these estimates capture both the direct benefit of a lower pupil-teacher ratio and any offsetting effects resulting from the expansion of the teaching force, the latter of which almost certainly depends upon labor market factors specific to the place and time period.

In contrast to the earnings papers that relate school inputs in childhood to earnings as an adult, the three achievement analyses investigate the effects of class size on end of year achievement. In order to account for unobserved school and neighbourhood factors and student differences, Rivkin, Hanushek, and Kain (2005) model test score gain as a function of class size, teacher experience and education, family characteristics, and full sets of student, school by year, and in some cases school by grade fixed effects. Equations 63 and 64 divide this model in two in order to highlight the school fixed effects:

$$A_{iGst} - A_{iG-1,s',t-1} = SC_{Gst}\beta_{gain} + \alpha_i + e_{iGst} \tag{63}$$

$$e_{iGst} = \omega_s + \xi_G + \psi_t + \rho_{Gt} + \pi_{sG} + \varphi_{st} + \tau_{sGt} + \varepsilon_{iGst} \tag{64}$$

[56]Theil (1954) examines aggregation in a linear framework.

where SC is a vector of the school characteristics and α is a student fixed effect. Equation 64 decomposes the error term into a number of school, grade, and year components and a random error. The first three terms are fixed school (ω), grade (ξ), and year (ψ) effects, the next three terms (ρ,π, φ) are second level interactions among these three components, the seventh term (τ) is the third level interaction, and the final term (ε) is a random error.

The school fixed effect (ω) captures time invariant differences in neighbourhoods and schools, many of which are likely related to both achievement and school racial composition. These include school facilities, public services, community type, and working conditions that influence teacher supply. The grade, year, and year-by-grade fixed effects (ξ, ψ, ρ) account for statewide trends in class size and achievement by grade and year and other factors including changes in test difficulty.

Because school quality may vary over time and by grade for each school, Equation 64 also includes interactions between school and both grade and year. The school-by-grade component (π) captures any systematic differences across grades in a school that are common to all years, and the school-by-year (φ) term accounts for systematic year-to-year differences that are common to all grades in a school. The school-by-grade fixed effects account for school or district specific influences on the quality of instruction that might vary by grade such as curriculum or information technology.

The school-by-year fixed effects remove in a very general way not only school specific performance trends but also idiosyncratic variation over time in school administration and in neighbourhood and local economic conditions that likely affect mobility patterns including such things as the introduction of school policies or local economic or social shocks. For example, an economic shock that reduces neighbourhood employment and income is absorbed and will not bias the estimates; nor will a shock to local school finances or the quality of the local school board, because each of these would affect all grades in a school.

The seventh term, τ, is the full three-way interaction between school, grade, and year; it cannot be included in the estimation, because there would be no variation left in class size across time or grades. Ignoring this three-way interaction means that grade specific variation over time in school average teacher quality or other achievement determinants could potentially bias the estimates if also correlated with class size. Yet the non-trivial costs of switching schools, the presence of multiple children in a family, and the fact that teacher assignments and other relevant aspects of school decisions are typically not known until immediately prior to the beginning of school year reduces the likelihood that changes over time in school and teacher quality for specific grades are systematically linked with yearly changes in class size through parental behavioural responses.

In this framework, the remaining variation in class size comes from differences across classrooms at point in time and differences in the pattern of grade average class sizes experienced by adjacent cohorts in a school that come from changes in policy regarding the allocation of resources among grades, students movement among schools and natural demographic variations in cohort composition. In terms of differences across classrooms, the potential for non-random allocation of students such as the placement of more difficult to educate students in smaller classes raises concerns about the validity of such variation, and the use of grade average class size in this study avoids the introduction of selection bias from this channel. In terms of differences in the pattern of grade average class size among adjacent cohorts, an identifying assumption in a number of studies that make use of cohort differences is that either raw cohort differences or differences remaining following the removal of school specific trends over time are not correlated with confounding factors. This approach builds on the intuition that students close in age in the same school have many similar experiences including similar quality teachers. Therefore this structure identifies the ceteris paribus class size effect, holding constant other school factors.

Despite the multiple levels of school fixed effects, unobserved student heterogeneity might introduce bias, possibly through the linkages between academic preparation and either the number of new entrants or cohort size. The use of achievement gain as dependent variable and inclusion of student fixed effects should, however, account for differences related to both student movement among schools and enrolment differences among adjacent cohorts.

Although the fixed effect in gains specification accounts for primary confounding factors, the use of achievement gain as dependent variable in a student fixed effects model biases the coefficients away from zero as discussed above. Moreover, the multiple fixed effects likely exacerbate any error in the measurement of class size, and the estimates are sensitive to the elimination of observations with class size values that appear to be incorrect. Finally, the exogeneity assumption that the remaining errors are orthogonal to class size may be violated if within school differences in class size across cohorts are related to unobserved differences in teacher quality or time varying student factors that affect achievement.

Angrist and Lavy (1999) use a regression discontinuity, instrumental variables approach based on Maimonides Law to identify class size effects on 4th and 5th grade achievement in Israel. Regression Discontinuity (RD) is a quasi-experimental method that uses a discontinuity in the probability of treatment to identify the local average treatment effect (LATE) and avoid bias introduced by non-random selection into treatment. Hahn, Todd, and Van der Klaauw (2001) describe identification conditions and estimation using an RD design, and we review their work prior to discussing the Angrist and Lavy estimates of class size effects.

Equation 65 presents a simple model of the effect of treatment x on outcome y for individual i:

$$y_i = \alpha_i + x_i\beta_i$$

$$\alpha_i \equiv y_{0i} \tag{65}$$

$$\beta_i \equiv y_{1i} - y_{0i}$$

where y_{0i} is the outcome without treatment and y_{1i} is the outcome with treatment. Hahn, Todd and Van der Klaauw begin by considering the case of a homogeneous treatment effect where $\beta_i = \beta$. Let z take on a continuum of values where the conditional probability

$$f(z) = E[x_i|z_i = z] = Pr[x_i = 1|z_i = z] \tag{66}$$

is discontinuous at $z_i = z_0$. Note that this is commonly referred to as a fuzzy RD design where other unobserved variables also affect the probability of treatment; the sharp design can be treated as a special case in which assignment to treatment is a deterministic function of z.

The key assumption required for identification is that $E[\alpha_i|z_i = z]$, is continuous in z at z_0, which is justified by the belief that persons close to the threshold are similar. The authors prove that β is non-parametrically identified as long as this assumption holds, the positive and negative limits for the probability of treatment exist at z_0, and the probability of treatment is discontinuous at z_0.

Hahn, Todd and Van der Klaauw turn next to identification when treatment effects are heterogeneous. They first prove that the local average treatment effect at z_0 is non-parametrically identified if the assumptions from the constant treatment effect case outlined in the previous paragraph are satisfied, the average treatment effect at z_i, $E[\beta_i|z_i = z]$, is continuous at z_0, and x_i is independent of β_i conditional on z_i near z_0.

The authors point out that the assumption that x_i is independent of β_i conditional on

z_i near z_0 assumes that the anticipated gains from treatment do not affect the probability of receiving treatment, a strong assumption that may well be violated in practice. Therefore the authors, drawing from Imbens and Angrist (1994), establish the identification of the local average treatment effect at z_0 under an alternative set of conditions that allows selection into treatment on the basis of prospective gains without the strong assumption of conditional independence. Specifically, they consider the case where treatment assignment is a deterministic function of z for each observation i, but the function is different for different groups or persons. Given this supposition, Hahn, Todd, and Van der Klaauw describe the specific assumptions necessary for identification of the local average treatment effect at z_0.

The authors then turn to estimation and propose the use of local linear nonparametric regression (LLR) methods rather than standard kernel estimators based on work by Fan (1992) showing that the LLR estimator has better boundary properties than the standard kernel estimator. They derive the asymptotic distribution of the RD treatment effect estimator based on LRR in the Appendix to the paper.

We now turn back to the class size application of RD based on Maimonides Law. This rule prohibits class sizes larger than forty, meaning that if schools desire to have class size of at least forty an increase in enrolment from 40 to 41 reduces average class size from 40 to 20.5, and increase in enrolment from 80 to 81 reduces average class size from 40 to 27, and so on. The authors argue that the use of predicted class size based on the rule as an instrument for class size along with flexible controls for enrolment effects produces consistent estimates of class size effects on 4th and 5th grade achievement, and the pattern of observed class sizes largely corresponds to that which would be predicted by Maimonides Law. As is the case with the study by Rivkin, Hanushek and Kain (2005), these estimates are aimed at capturing the pure effect of smaller classes holding all other factors constant. In this case the estimates can be interpreted as weighted averages of local average treatment effects across the various boundaries,

where the weights reflect the numbers of schools that contribute to identification by having enrolment that places it at a particular boundary.

There are reasons to be concerned that the identification conditions described in Hahn, Todd and Van der Klaauw (2001) could be violated. First, the paper shows that many schools add classes prior to enrolment reaching a multiple of forty, and districts may manipulate enrolment among schools in order to comply with the rule. Consequently the assumption that expected achievement in the absence of the small class is continuous at the boundary may be violated, as schools with enrolments just above boundaries may differ systematically from those just below. Moreover, the basic models combine the effects of changes in class size around the boundaries and intra-boundary changes, and intra-boundary class size variation may well be correlated with unobserved determinants of achievement. Restricting identification to comparisons across boundaries by limiting the sample to schools with enrolment that is first within five students and then within three students of a boundary leads to fluctuation in the estimates, and the increasingly small samples also reduce precision. Moreover, such selection may be endogenous because of the way the system is actually administered in practice.

In contrast to the other four papers based on observational data, Krueger (1999) uses data generated by a random assignment experiment designed to uncover the benefits of smaller classes. In an ideal experiment in which both students and teachers were randomly assigned to class types, mean achievement comparisons between treatment and control groups would produce unbiased estimates, and family background, teacher, and school information could be included to reduce sampling error. However, non-random movement between treatment and control groups and non-random attrition from the sample potentially introduces selection bias. Although the use of initial random assignment as an instrument for actual class type can produce LATE estimates, subject to the monotonicity assumption, selective attrition provides a more vexing

problem.

Similar to the two other class size papers, the Tennessee STAR experiment is designed to produce estimates of the direct benefit of smaller classes ignoring any change in the quality of instruction. This potentially diverges from the benefit that would be realised in a large-scale class size reduction that requires substantial expansion of the teaching force.[57]

5.4 Estimation of Teacher Value-Added

The passage of No Child Left Behind its requirement that states test students annually and build comprehensive data systems has expanded opportunities to estimate teacher productivity as measured by value-added to student achievement. The repeated test scores and tracking of students and teachers through time enable researchers to account for differences among students and schools that could impede efforts to identify teacher value-added. In this section we examine methods used to estimate teacher value added in five papers listed in Table 3. The first four estimate effectiveness for each teacher using regression models that account for potential confounding factors in different ways, while the final paper estimates the variance in teacher quality on the basis of the pattern of school average achievement.

Despite the steps taken to account for confounding factors in these and other papers, Rothstein (2009) and others have begun to raise concerns about the methods used to measure teacher quality. These critiques argue that sorting on unobservables that vary over time, endogenous parental response to teacher quality, test measurement error, and other failings introduce bias to estimates of teacher value added and estimates of the variance in teacher value added.

[57] Jepsen and Rivkin (2009) estimate the effects of changes in teacher experience and certification that accompanied class size reduction in California, but a lack of data impedes efforts to learn more about the magnitude, timing, and distribution of any decline in teacher quality.

Table 3: Selected Research on Variance in Teacher Quality

Paper	Method	Data	Findings
Aaronsen, Barrow, and Sander (2007)	value-added model; teacher fixed effects; school fixed effects in some models; observed student characteristics	Chicago Public Schools administrative data	significant variation in teacher quality including variation by experience
Rockoff (2004)	value-added model; teacher fixed effects; student and school by year fixed effects; observed student characteristics (check)	New Jersey school district administrative data	significant variation in teacher quality, conditional on experience
Ballou, Sanders, and Wright (2004)	first estimate separate teacher specific residual test score gains for each subject and grade. Then produce a single quality estimate for each teacher based on the variance-covariance structure of these residual gains.	administrative data	significant variation in teacher quality including variation by experience (check)
Kane and Staiger (2008)	regress difference in average test scores for two teachers in a school randomly assigned to classrooms on earlier year difference in Empirical Bayes estimates of value-added for the same pair.	experimental and non-experimental data for a small number of Los Angeles public schools	significant variation in teacher quality in non-experimental and experimental data; do not reject the hypothesis that non-experimental estimates are unbiased.
Rivkin, Hanushek, and Kain (2005)	compare cohort differences in test score gains with share of teachers who are different in the respective cohorts in a school.	Texas public school administrative data	significant variation in teacher quality within schools

In order to highlight the issues of bias and sampling error that are addressed in each of these papers, Equation67 decomposes the estimate of teacher value-added for teacher j in year y as the sum of the true teacher effect (assumed not to vary over time), the confounding student contribution (subscript i), the confounding peer contribution (subscript p), the confounding school contribution (subscript s), and random sampling error (subscript n):

$$\hat{t}_{jy} = t_j + \hat{\varepsilon}_{iy} + \hat{\varepsilon}_{py} + \hat{\varepsilon}_{sy} + \hat{\varepsilon}_{ny} \tag{67}$$

Estimates of teacher value added deviate from the true teacher effect, but if the expected values of each of the four error terms are zero unobserved differences in student, peer, and school characteristics would not introduce bias. Regardless, the variance of $\hat{t}_{jy}$ incorporates the true variance in teacher quality plus the variances of the other terms. Thus estimation of the variance in teacher value added must address complications related to both bias and sampling error, and the methods frame the interpretation of the estimates.

Aaronson, Barrow and Sander (2007) use lagged achievement and observed characteristics to account for unobserved student heterogeneity and school fixed effects to control for school and peer differences that would otherwise be captured by the teacher effects; experience controls are not included, meaning that the estimated effects combine fixed differences across teachers and differences related to experience. The school fixed effects are omitted from some models, because in addition to accounting for confounding school factors they also soak up any systematic sorting by quality of teachers into schools. Importantly, the school fixed effects do not mitigate bias resulting from sorting into classrooms on the basis of unobserved time varying or even fixed differences in the rate of learning not captured by the included lagged achievement measures.

Rockoff (2004) takes a different approach to accounting for unobserved heterogeneity; he includes student fixed effects but not measures of prior achievement, implicitly

imposing the strong assumption of no knowledge depreciation over time. He also includes school by year fixed effects to eliminate any between school variation including systematic differences in teacher quality. Finally, he controls for teacher experience in order to isolate fixed differences in teacher effectiveness.

Ballou, Sanders, and Wright (2004) use a sequential process to estimate teacher effects purged of the influence of student heterogeneity. First, they regress test score gain on a vector of student characteristics and teacher fixed effects separately by grade and subject. Rather than simply treating the estimated teacher fixed effects as estimates of teacher quality in a particular subject and grade, they combine information from different subjects and grades in order to produce a single quality estimate for each teacher. This is accomplished in the following steps: 1) use the student demographic variable coefficients obtained from the teacher fixed effect models to subtract the contributions of the student variables from test score gain; and 2) use the variance/covariance structure of teacher average residual test score gains for all grades and subjects to produce a single quality estimate for each teacher. Note that although the use of teacher fixed effects in the first stage eliminates bias introduced by sorting into classrooms on the basis of the included variables, omitted student factors introduce bias if students sort into schools or classrooms on the basis of unobserved factors.

The authors claim that the insensitivity of the estimates to the inclusion of the student covariates provides evidence that the use of multiple tests accounts for the confounding effects of unobserved heterogeneity, but this finding is not surprising given that the limited set of covariates explains little of the achievement variation within classrooms, accounts for little of the heterogeneity among students, and may well be unrelated to unobserved confounding variables. The strong implicit assumptions about the nature of sorting among schools and classrooms, about the contributions of school and peer effects, about the covariance of teacher effectiveness across subjects and years, and about manner through which knowledge accumulates through time are unlikely to

be satisfied, and this may introduce substantial bias. The fact that the estimates were far more sensitive to the introduction of peer characteristics than to student level controls suggests that unobserved school differences and systematic student sorting by school may present particular problems.

Recognising the threat of student sorting both within and between schools to the estimation of teacher value-added, Kane and Staiger (2008) use experimental data generated by a random assignment study of the National Board for Professional Teaching Standards Certification Program to investigate the validity of non-experimental estimates of teacher value-added. In the study, pairs of teachers are identified in each school, one with and the other without certification, and classrooms are randomly assigned to the pairs. The difference in average test scores of the classrooms is regressed on the difference in Empirical Bayes estimates of value added for the pair of teachers based on multiple years of data from earlier years in order to examine the validity of the estimation based on non-experimental data.[58] The hypothesis test is based on the estimate of β in the regression of average achievement in teacher j's classroom on VA, the empirical Bayes value-added estimate for teacher j:

$$\bar{A} = \beta VA_j + \varepsilon_p \tag{68}$$

It is the structure of the empirical Bayes estimator that underlies the hypothesis test. Specifically, VA_j equals the random effect estimate for teacher j multiplied by

$$a = \frac{\sigma_t^2 + \sigma_{\bar{i}}^2 + \sigma_{\bar{p}}^2 + \sigma_{\bar{s}}^2 + 2\sigma_{t,\bar{i}}^2 + 2\sigma_{t,\bar{p}}^2 + 2\sigma_{t,\bar{s}}^2}{\sigma_t^2 + \sigma_i^2 + \sigma_p^2 + \sigma_s^2 + 2\sigma_{t,i}^2 + 2\sigma_{t,p}^2 + 2\sigma_{t,s}^2} \tag{69}$$

Note that the bars in the numerator indicate that the terms capture the persistent components of individual, peer, and school variation among teachers.[59] Thus the magnitude of a determines the extent to which the estimate for teacher j is shrunk toward

[58] See Morris (1983) for a discussion of the empirical Bayes estimator.

[59] To simplify we set the covariances among the individual, peer, and school factors equal to zero.

the grand mean teacher quality of zero: the lower the ratio of the persistent components to the total variance the more the estimate is shrunk toward zero.

The expected value of β equals

$$\begin{aligned} \frac{cov(\bar{A}_j, VA_j)}{var(VA_j)} &= \frac{\sigma_t^2}{a\left(\sigma_t^2+\sigma_i^2+\sigma_p^2+\sigma_s^2+2\sigma_{t,i}^2+2\sigma_{t,p}^2+2\sigma_{t,s}^2\right)} \\ &= \frac{\sigma_t^2}{\left(\sigma_t^2+\sigma_{\bar{i}}^2+\sigma_{\bar{p}}^2+\sigma_{\bar{s}}^2+2\sigma_{t,\bar{i}}^2+2\sigma_{t,\bar{p}}^2+2\sigma_{t,\bar{s}}^2\right)} \end{aligned} \tag{70}$$

If there were no persistent differences in student, peer, or school components across teachers not accounted for in the model, then all terms following σ_t^2 would equal 0, and the ratio would equal 1. This suggests a test of the null hypothesis of β=1 as a specification test: rejection of the null hypothesis would provide evidence in support of the presence of non-random sorting on unobservables. Kain and Staiger report estimates that range from roughly 0.75 to 1.1 in their preferred specifications that control for student heterogeneity with lagged test scores. Importantly, none of these estimates are significantly different from one, which is consistent with the hypothesis that sorting on unobservables does not confound the estimates of teacher value-added based on observational data.

It should be noted that the test does have some limitations. First, given the small sample size, even if the 95 or even 90 percent confidence interval for β contains 1, it also contains values that are much smaller than one that would be evidence of sorting on unobservables. Second, if there is compensatory assignment of better teachers to more difficult students, the covariance terms would be negative and would offset some of the persistent variation in student, school, or peer differences among teachers not captured by the model, potentially pushing the estimate toward 1. Finally, the small group of schools in which principals agreed to permit classes to be randomly assigned to teachers is unlikely to be representative, meaning that evidence of the validity of

value-added estimates with this sample may not generalise beyond this sample.

The final paper by Rivkin, Hanushek, and Kain (2005) avoids the question of sorting within classrooms altogether by focusing on cohort differences in achievement gains within schools. Specifically, the approach builds on the notion that if schools select teachers from a pool with substantial variation in quality, higher teacher turnover should lead to larger differences in test score gains between adjacent cohorts as fewer students in a cohort have a teacher who also taught the other cohort. Equation 71 represents average achievement gain in grade g in school s for cohort c as an additive function of grade average student and teacher fixed effects, a school fixed effect and the grade average error:

$$\Delta\bar{A}^{c}_{gs} = \bar{\gamma}^{c}_{gs} + \bar{\theta}^{c}_{gs} + \delta_{s} + \bar{\upsilon}^{c}_{gs} \tag{71}$$

Taking the difference between adjacent cohorts c and c' in the differences of grade average gains in achievement in grades g and g-1 for the sample of students who remain in the school in both grades eliminates all fixed student and family differences, leaving only cohort-to-cohort differences in the grade average difference in teacher quality and time varying student and school factors (contained in υ) as determinants of the difference in the pattern of achievement gains.

$$\begin{aligned} &\left(\Delta\bar{A}^{c}_{gs} - \Delta\bar{A}^{c}_{g's}\right) - \left(\Delta\bar{A}^{c'}_{gs} - \Delta\bar{A}^{c'}_{g's}\right) = \\ &= \left[\left(\bar{\theta}^{c}_{gs} - \bar{\theta}^{c}_{g's}\right) - \left(\bar{\theta}^{c'}_{gs} - \bar{\theta}^{c'}_{g's}\right)\right] + \left[\left(\bar{\upsilon}^{c}_{gs} - \bar{\upsilon}^{c}_{g's}\right) - \left(\bar{\upsilon}^{c'}_{gs} - \bar{\upsilon}^{c'}_{g's}\right)\right] \end{aligned} \tag{72}$$

Squaring this difference yields a natural characterisation of the observed achievement differences between cohorts as a series of terms that reflect variances and covariances of the separate teacher effects plus a catchall component e that includes all random error and cross product terms between teacher and other grade specific effects.

$$\left[\left(\Delta\bar{A}^{c}_{gs}-\Delta\bar{A}^{c}_{g's}\right)-\left(\Delta\bar{A}^{c'}_{gs}-\Delta\bar{A}^{c'}_{g's}\right)\right]^2 = \quad (\bar{\theta}^{c}_{gs})^2+\left(\bar{\theta}^{c}_{g's}\right)^2+\left(\bar{\theta}^{c'}_{gs}\right)^2+\left(\bar{\theta}^{c'}_{g's}\right)^2$$

$$-2\left(\theta^{c}_{gs}\bar{\theta}^{c'}_{gs}+\theta^{c}_{g's}\bar{\theta}^{c'}_{g's}\right)$$

$$+2\left[\left(\theta^{c}_{gs}\bar{\theta}^{c'}_{g's}-\theta^{c}_{gs}\bar{\theta}^{c}_{g's}\right)+\left(\theta^{c}_{gs}\bar{\theta}^{c'}_{g's}-\theta^{c'}_{gs}\bar{\theta}^{c'}_{g's}\right)\right]+e \tag{73}$$

Under assumptions that formally characterise the notion that teachers are drawn from common distributions over the restricted time period of the cohort and grade observations, the expectation of Equation 73 yields:

$$E\left[\left(\Delta\bar{A}^{c}_{gs}-\Delta\bar{A}^{c}_{g's}\right)-\left(\Delta\bar{A}^{c'}_{gs}-\Delta\bar{A}^{c'}_{g's}\right)\right]^2 = 4\left(\sigma^2_{\bar{\theta}_s}-\sigma^2_{\theta^{c}_{s}\theta^{c'}_{s}}\right)+E(e_s) \tag{74}$$

where $\sigma^2_{\bar{\theta}_s}$ is the variance of teacher quality in school s and $\sigma^2_{\theta^{c}_{s}\theta^{c'}_{s}}$ is the covariance in teacher quality across cohorts in a school.

Equation (74) provides the basis for estimation of the within-school variance of teacher quality over the sample of students that remain in the same school for both grades. The left-hand side is the squared divergence of the grade pattern in gains across cohorts, which is regressed on the proportion of teachers in a school who are different in cohort c' than in cohort c. In order to account for differences in the number of teachers and place all schools on a common metric, the proportion different must be divided by the number of teachers per grade, and the coefficient on this proportion divided by four provides the estimate of the within-school variance in teacher quality. Only unobserved, time-varying factors systematically related to teacher turnover can introduce bias, and sensitivity testing suggests that any such biases are negligible.

As previously noted, the use of test score gain likely introduces some upward bias,

though violation of the strong assumption that true teacher effectiveness never varies over time and errors in the measurement of both the number of teachers in a grade and turnover both bias the estimates toward zero. Moreover, the estimates based on this approach ignore all between school differences in the quality of instruction. Finally, a limitation of this aggregate approach for policy is the absence of estimates of value-added estimates for individual teachers.

Finally the issue of scale for the test scores reappears in this literature. Many studies of teacher value added rely on a specific test score scale and are based on comparing gains, which are not invariant to monotonic transformations. Moreover, the empirical strategy of differencing out heterogeneity relies on the assumption of linearity for the particular score at hand. This issue raises questions about the robustness of the results to changes of scale and merits attention during sensitivity testing given the absence of an agreed upon metric to anchor the results. Moreover, this concern supports the use of more flexible parameterisation of prior achievement as controls.

One source of bias for all approaches to the estimation of teacher value-added or the variance in teacher quality is the endogenous intervention of parents and schools. Todd and Wolpin (2003) and Dearden, Ferri, and Meghir (2002) discuss the likelihood that the amount of time and money dedicated to academic support is likely to depend on the quality of instruction. Cullen, Jacob, and Levitt (2006) find mixed evidence regarding the effect of winning a lottery to choice into a specific school on parent involvement, but individual teacher quality may induce a stronger parental response. Though inspiring teachers can potentially induce parents to become more involved, parental intervention to compensate for lower quality instruction is more likely. In addition to influencing parental behaviour, teacher quality may also affect the amount of intervention support allocated by the school. For example, reading or mathematics specialists may spend additional time in classrooms with less effective teachers.

Such compensatory intervention by school staff and parents would tend to bias

value-added estimates toward the school mean and estimates of the within school variance in teacher quality toward zero due to the negative correlation between teacher quality on the one hand and both the school and parent components on the other. Even the random assignment of classrooms to teachers does not mitigate the impact of this type of endogenous response to realised quality, and any such biases would not be detected in the specification test proposed by Kane and Staiger (2008). Therefore in the absence of controls for parental and school interventions such as the quantity and quality of family and school support in specific subjects, teacher value-added estimates capture both classroom teacher effects and the contributions of other sources of academic support.

5.5 Estimation of the housing market capitalisation of school quality

The belief that the quality and cost-effectiveness of local public schools affect housing values provides a key underpinning for the notion that competition among localities fosters higher quality public services; this issue is the focus of Chapter ? in this volume. Yet the non-random sorting of families into communities, multitude of public services provided, and difficulty controlling for all housing and neighbourhood amenities impede efforts to empirically test the relationship between housing prices and school quality. Recent work has attempted to overcome these difficulties by focusing on comparisons of houses on opposite sides of school attendance zone boundaries. The validity of school quality capitalisation models with boundary fixed effects rests in large part on the assumption that unobserved determinants of housing prices vary continuously at the boundary and are virtually uncorrelated with school quality differences between houses on opposite sides of the boundary. This is the continuity assumption described in Hahn, Todd, and Van der Klaauw (2001). If it is satisfied, differences

in school quality would account for any discontinuity in housing prices at the school attendance zone or school district boundary, and boundary fixed effects models would generate consistent estimates of the relationship between measured school quality and price at the boundary.

Table 4 lists three papers that utilise somewhat different boundary fixed effect models. Despite their differences, each finds that accounting for unobserved neighbourhood differences substantially reduces the estimated relationship between house price and school average test score. Although remaining concerns about bias introduced by unobserved differences in housing quality, the characteristics of immediate neighbours or the quality of other amenities remains an important issue, a key methodological question is what exactly underlies any relationship between the measures of school quality and the house price.

Black (1999) calculates differences in mean house prices on opposite sides of school attendance zone boundaries and investigates whether the differences, are systematically related to the differences in test scores in the respective schools, adjusted for a set of observed housing characteristics. Only boundaries in which both attendance zones lie in the same city and school district are included, so this method holds constant the property tax rate, district administration, the quality of city public services, and other amenities that do not vary within the narrow boundaries. The focus on close neighbours also has the advantage of accounting for the effects of factors that change over space such as proximity to parks, police and fire stations, and public transportation.

One potential threat to the identification of the capitalised value of higher test scores is the possibility that the parsimonious set of housing variables fail to capture quality differences that may be related to test scores; higher income families may both select the house in the higher test score zone and spend more on home renovation, introducing an upward bias in the estimate of the capitalised value of higher scores. Any failure

Table 4: Selected Research on Boundary Fixed Effect Estimates of Housing Market Capitalisation of School Quality

Paper	Comparison	Method	Data	Findings
Black (1999)	attendance zone boundaries within school districts	relate differences in school average test score to house price differences across boundaries, controlling for observables	house price data and information on school achievement for districts in Massachusetts	evidence of housing market capitalisation of test scores
Bayer, Ferreira, and McMillan (2007)	attendance zone boundaries within school districts	relate differences in school average test score to house price differences using reduced form and discreet choice boundary fixed effects models.	house price data, restricted US census demographic data and information on school achievement for districts in San Francisco Bay Area	inclusion of neighbourhood demographic characteristics reduces the estimated effect of test score. Evidence of heterogeneity in willingness to pay for school quality
Gibbons, Machin and Silva (2009)	local education authority boundaries and empirically determined attendance zone boundaries within local education authorities	relate differences in school average test score to house price differences across boundaries, controlling for observables	UK administrative data sources and census data	evidence of housing market capitalisation of school value-added and average achievement

of the included characteristics to capture dimensions of housing quality that are related to school average test score including the demographic characteristics of neighbours will lead to a non-zero within boundary covariance between the error and test score and introduce bias. The direction of the bias would likely be negative if unobserved housing quality were higher on the low test score side of the boundary, consistent with heterogeneous preferences regarding education, and would likely be positive if higher income households tended to live in the higher test score side of boundaries.

An important limitation of this approach is the inability to disentangle the contributions of peers from that of the quality of the provision of public education per tax dollar spent. The restriction that boundaries lay within as opposed to between school districts mitigates biases potentially introduced by confounding factors across district boundaries at a cost of eliminating any impact of district policies or resource use on house prices. As the district controls principal and staff hiring, curricular decisions, capital investments this is a major drawback, making it more likely that student demographic characteristics play a primary role in the determination of the desirability of an attendance zone.

Bayer, Ferreira, and McMillan (2007) embed a boundary fixed effects approach in a model of neighbourhood choice using restricted U.S. Census data that provides richer information on the characteristics of neighbours. The finding that the inclusion of neighbourhood socio-demographic characteristics roughly cuts in half the estimated effect of school average test score even in models with rich controls for average student characteristics suggests that some of what appears to be preference for school or peer quality is actually a preference for neighbour demographics. Importantly, average test score is likely to be a poor proxy for school effectiveness in raising achievement, and a more accurate measure of school value-added would likely provide a clearer picture of the value placed on more effective schools.

The difference in neighbourhood characteristics on opposite sides of the boundar-

ies suggests the existence of heterogeneous preferences regarding education, and this is precisely what is shown by the structural estimation of willingness to pay for school quality and other neighbourhood characteristics. The discrete choice model incorporates heterogeneity in the willingness to pay for higher test scores and other house characteristics and neighbourhood amenities. It should be noted that the focus on within district comparison of attendance zones prevents differences in district quality from being capitalised into higher house prices in both the hedonic price and discrete choice regressions.

The additional structure requires the satisfaction of a number of assumptions including 1) place of work is exogenous; 2) housing characteristics more than three miles from any house have no direct effect on the residents; 3) the included neighbourhood characteristics account fully for sorting across boundaries; 4) school average test scores and other school characteristics control fully for differences in school quality; and 5) the multinomial logit IIA assumptions. Concerns can be raised by each of these assumptions, but the underlying assumption of no variation in the weighting of different aspects of school quality merits additional attention in a framework that emphasises heterogeneity in preferences. The possibility of variation in the quality of instructing students from different points in the achievement distribution or different backgrounds cannot be dismissed, and this model may misinterpret that as differences in willingness to pay for an amenity whose quality does not vary by student or family characteristics.

Gibbons, Machin, and Silva (2009) use boundary fixed effect hedonic models to estimate capitalised values of both school effectiveness, as measured by test score gain, and school composition, as measured by initial test score, for a sample of United Kingdom students. In contrast to the two other papers, this paper focuses on boundaries between local authorities, which perform many of the same functions as US school districts. Therefore local authority actions that affect achievement gain are captured in the estimates. Moreover, test score gain would appear to provide a better measure of

school value-added than average test score.

Because of the absence of school attendance zones within local authorities, the authors had to construct a method for measuring the quality of schooling associated with each housing unit. They used the empirical distribution of school attendance to define a set of overlapping school attendance zone boundaries, meaning that many houses were located in more than one attendance zone. Subsequently school characteristics were computed for each house as a weighted average of the initial test score and test score gain for schools in all the relevant attendance zones.

A concern about this approach to measuring school quality is the fact that families are assumed to respond to the same test score and gain information that their actual location decisions help to determine (for example through peer effects). This problem holds for all the papers using test score information and relates to the aforementioned possibility that test scores capture differences in unobserved quality of neighbours or neighbourhood amenities. It is quite difficult to distinguish whether 1) test score differences reflect differences in school quality; 2) test score differences result from sorting on the basis of school reputations; and 3) test score differences result from sorting on the basis of other amenities including the characteristics of residents. In fact each of these channels could contribute to test score differences across boundaries.

The use of measures of school quality less directly related to family characteristics would mitigate this problem, but input measures historically explain little of the variation in school effectiveness. The estimation of school value added controlling for observed and unobserved student heterogeneity would provide a better measure of school quality with which to estimate the capitalisation of school effectiveness into housing prices.

5.6 Estimation of the Effects of Competition, Choice, and Accountability

Recent expansions in public school choice and the growth of accountability systems in many countries have altered the public school environment, but the effects of such programs have proven elusive to estimate because of both the manner in which the programs have been introduced and the difficulty accounting for student and family heterogeneity related to different choices. A number of methods have been used to identify choice and accountability effects, including instrumental variables, difference-in-differences and lottery outcomes as instruments for school enrolment. Table 5 provides examples of three of these methods used in studies of choice and accountability. The study of catholic-public school quality differences by Neal (1997), the study of accountability system effects on achievement by Hanushek and Raymond (2005), and the investigation of open enrolment effects by Cullen, Jacob and Levitt (2006) provide examples of these three methods.

The endogeneity of the decision to attend private school introduces a difference between students in public and private school that biases estimates of sector differences in quality unless the unobserved heterogeneity is fully accounted for. Consider the following two equation model. In the first equation achievement A is a function of a school sector indicator variable P (equal to 1 for catholic school and 0 for public school), a vector of family and community characteristics X, and an error u. In the second equation the unobserved propensity to attend private school $P*$ is a function of a vector of family and community characteristics Z and an error e. If $P* > 1$ students attend catholic school and $P = 1$; otherwise students attend public school and P=0.

$$\begin{aligned} A_i &= P_i\delta + X_i\beta + u_i \\ P_i^* &= Z_i\gamma + e_i \end{aligned} \tag{75}$$

Table 5: Selected Research on Choice and Accountability

Paper	Method	Data	Findings
Hanushek and Raymond (2005)	difference-in-differences	National Assessment of Educational Progress	accountability increases achievement
Neal (1997)	estimate catholic school effects using catholic church adherents as a share of the county population and number of catholic secondary schools per square mile as instruments	NLSY	effects on educational attainment from catholic school attendance are significant for urban minorities, modest for urban whites, and negligible for suburban students.
Cullen, Jacob and Levitt (2006)	use lotteries for over-subscribed schools to identify benefits of open enrolment in Chicago public schools	Chicago Public School District administrative data	attendance at a non-neighbourhood school does not significantly increase achievement

OLS estimation of the top equation is likely to produce an upward biased estimate of δ and overstate the catholic school-public school quality differential because the expected value of the error u is likely to be higher for students attending private school. A solution to this problem is the identification of an instrument that belongs in Z but not X, i.e. affects the probability of attending private school but otherwise does not affect achievement.[60]

Neal (1997) argues that catholic church adherents as a share of the county population and the number of catholic secondary schools per square mile provide valid instruments for the estimation of δ, because they affect the money (first instrument) or time (second instrument) costs of attending catholic school but are otherwise unrelated to A. Although religious affiliation may be a stronger predictor of catholic school attendance, Neal argues that it is not a valid instrument because of the possibility that it is directly related to achievement. For example, families may choose to become catholic in order to send a child to a catholic school.

The argument that personal religious affiliation is likely related to confounding factors while the county mean share of individuals with a particular religious affiliation is not related to those factors appears tenuous, as it seems likely that students in more heavily catholic counties with more catholic schools per square mile are more likely to have family backgrounds and even tastes similar to the average catholic. This would introduce a non-zero correlation between these instruments and u and bias estimates of δ. Moreover, the share of Catholics in the county population and number of catholic schools may be directly related to the quality of public schools by affecting both the level of tax support for the public schools and the involvement of community members in public school affairs. Such a correlation would also introduce an upward bias into the estimate catholic-public school gap.

Finally as Neal points out, this approach relies on the assumption that the residential choice is exogenous with respect to the quality of public and catholic schools. Given

[60] See Heckman (1979) for a comprehensive treatment of non-random selection.

the previously discussed evidence that families appear to sort partly on the basis of preferences regarding education, this assumption may well be violated. If the choice is between living in School District A and attending catholic school and living in School District B and attending public school, county composition is likely to be related to the quality of the public schools. The fact that the instrument is measured at the county level and inclusion of a number of demographic characteristics likely mitigate the potentially problematic impact of sorting, though many metropolitan areas include multiple counties and demographic variables may not capture salient differences among families and communities.

In terms of interpretation, this IV approach rules out consideration of general equilibrium effects of the presence of catholic schools affecting the quality of local public schools either through competitive pressures or financial support. Rather it identifies the race specific average catholic-public school difference in school value-added based on county differences in catholic religiosity. Given housing constraints faced by blacks in many metropolitan areas during the late 1970s, these arguments justifying the empirical approach seem stronger for blacks than for whites.

Cullen, Jacob and Levitt (2006) investigate choice entirely within a single public school district (open enrolment) through comparisons of lottery winners and losers. The use of lottery outcomes circumvents problems introduced by the fact that those who apply to non-neighbourhood schools differ from non-applicants and winners who decide to accept admission differ from those who decide not to accept admission.

Estimation of the following OLS specification generates estimates of the benefits of open enrolment:

$$A_{ia} = Win\_Lottery_{ia}\delta + Lottery_a\Gamma + +X_i\beta + u_{ia} \tag{76}$$

where Win_Lottery is a dummy variable equal to 1 if application a for student i was a lottery winner and Lottery is vector of lottery fixed effects that indicates to which

lottery the observation refers (there are 194 different lotteries, and students may participate in more than one). In this model the estimate of δ is the weighted average of the regression adjusted difference in mean outcomes for winners and losers of the various lotteries. In the absence of selective attrition from the sample or contamination of the lotteries the lotteries produce an unbiased estimate of δ.

The possibility of participating in multiple lotteries and not attend a lottery school even if you win the lottery affects the interpretation of δ. Choosing to participate in more lotteries raises the probability of winning and the probability of not attending a particular lottery school even if you win that lottery. If the treatment is defined as the average effect of attending the lottery school, the estimate of δ captures the intention to treat effect on applicants. The authors also provide an alternative explanation: the average impact of having a school in the choice set for students who expressed an interest. Note that as the number of choices and lotteries per-student rise, one would expect a decrease in the benefit to winning as fewer students would take up the opportunity to attend the lottery school and the next-best alternative for losers would tend to be closer in expected match quality to the lottery school.

Importantly, this and other lottery based analyses provide no information on either the general equilibrium effects of choice on the overall distribution of school quality or the specific factors that account for any positive or negative effects. Because the lottery analyses compare the outcomes of winners and losers, the estimates ignore any district wide increases or decreases in school quality. In terms of the sources of any differences between those attending non-neighbourhood and neighbourhood schools, the lotteries do not provide information on the contributions of teachers, facilities, curriculum, or peers.

The limitations of lottery based analyses in terms of public policy are perhaps most severe in cases where the set of lottery school spaces is small relative to the total number of students. The fact that participants in such lotteries are a selective sample means

that students in schools made up entirely of lottery winners will also constitute a selective group. Consequently a finding that winners of lotteries to charter or private schools outperform the losers who end up primarily at neighbourhood schools could be driven entirely by differences in the peer composition. If that is the case, program expansion would tend to diminish the average benefits to lottery winners as the lottery sample became less select, and the benefits of attending a charter or private school would decline to zero if all students were to attend such schools.

Hanushek and Raymond (2005) use a difference in differences framework to analyse the effect of accountability on the level and distribution of achievement. In contrast to school choice analyses where endogenous decisions on the part of families complicate the estimation, empirical studies of accountability effects must address issues related to the political decision to adopt an accountability regime at a particular time. Adoption affects all districts in a state, and it is likely that adoption is related to other contemporaneous policy changes such as a decision to add resources that also affect outcomes. Identification of accountability effects requires the construction of valid counterfactuals for accountability regimes.

Equation 77 describes state average achievement (A) in year y as a function of an indicator for an accountability regime C, a vector of a characteristics X that vary by state and year, and a composite error that includes a year effect α, a state effect τ and random term ε that varies by state and year.

$$A_{sy} = C_{sy}\delta + X_{sy}\beta + \alpha_y + \tau_s + \varepsilon_{sy} \tag{77}$$

Consider the following four possible estimators of the accountability effect. The first uses cross-sectional data for a single year, and identifies the accountability effect as the mean achievement difference between states with and without accountability systems. Clearly myriad differences among states would introduce a correlation between C and τ and contaminate the estimate.

The second uses time-series data for a single state that implemented an accountability system during the sample period. This method identifies the accountability effect as the mean achievement difference between periods following adoption and periods prior to adoption. In this case, any changes over time would confound the estimated accountability effect by introducing a correlation between C and α.

The third method uses panel data on states and a state and year fixed effects specification with a dummy variable for whether or not a state has an accountability program in year y. The fixed effects explicitly account for differences among states common to all years τ and all differences among years common to all states α. Identification relies on the assumption that changes over time in states that do not change accountability status provide a valid counterfactual estimate of what would have occurred in states that transition to an accountability system. This assumption that the covariance between C and ε equals zero would be violated if adoption of an accountability system were correlated with other changes that affect educational outcomes, perhaps including changes in the political system, economic circumstances, etc.

The fourth method used by Hanushek and Raymond adds a state specific time trend to account for state specific changes over time that could bias estimates of δ; the paper also measures accountability by the share of the previous four years covered by an accountability system. In this framework the accountability effect is identified by within-state deviations from the time trend related to the adoption of an accountability system and differences in the timing of accountability adoption across states. Although the addition of a state-specific time trend controls for trends over time that could confound the estimates, it does not account for discontinuous time-varying factors include those related to the decision to enact an accountability structure. Even with long time series and polynomial trends the possibility remains that accountability program adoption would not be the only factor contributing to a discontinuous change in achievement around the time of program adoption.

Interpretation is complicated by a number of factors including heterogeneity in accountability programs, uncertainties in the time pattern of effects following program adoption, and the extent to which the adoption of accountability programs led to changes in other factors that affect achievement including the amount of resources devoted to education. Hanushek and Raymond control for school spending, but to the extent that accountability increases the return on investments in education, the magnitude of the effect varies with school spending. Therefore a specification that restricts the effect to be constant when it is in fact inter-woven with the level of spending will provide an incomplete picture of the overall effects. In addition, the long run effects may include any impact on the quality and distribution of the stock of teachers, while effects in the short run operate largely through other mechanisms.

6 Conclusions

Estimation of the return to schooling or school input effects must account for the complications introduced by the myriad choices made by families, teachers, administrators, politicians, judges, or other actors. Approaches to estimation range from structural models derived explicitly from theory to experimental and quasi-experimental methods based on randomised trials, rule changes, policy changes, lotteries or other source of variation. Although these methods may appear as almost polar opposites in terms of their reliance on theory and explicit assumptions about behaviour, closer inspection says otherwise. Rather conditions for identification and parameter interpretation in most cases require assumptions about some aspects of the underlying choice framework. Importantly, such assumptions appear to take on greater importance as the models become more flexible and comprehensive. As we highlight in Part 1, identification of the return to schooling in models that allow for heterogeneous returns and multiple dimensions of unobserved heterogeneity require assumptions about the processes

underlying the choice of schooling level. Similarly, efforts to estimate differences in teacher quality require assumptions regarding the mechanisms through which teachers and students are matched. In addition, they also require assumptions about the behaviour of parents in response to observed instructional quality in order to identify teacher effects on learning. The proliferation of elementary and secondary school administrative data facilitates the use of panel data methods that can account for confounding factors introduced by purposeful choices. Moreover, the linking of these data with information on wages, involvement in the criminal justice system, and vital statistics will enable researchers to follow children from birth to adulthood. This expands the range of questions that can be addressed but also introduces additional choices and behaviours that must be accounted for. More complicated conceptual and empirical models will be required to identify treatment effects in these settings. Although there are many directions to expand and improve upon existing work, four specific areas come to mind based upon existing findings and recent methodological developments. The first is improved treatment of individual heterogeneity along multiple dimensions including non-cognitive skills; the second is the incorporation of endogenous parental responses into estimates of school input effects or teacher quality; the third is an enhanced understanding of the linkages among school and teacher effects across years and classrooms including the effects of heterogeneity in preparation on the distribution of teacher effort and learning; and the fourth is a greater integration of differences in school quality into models of education choice and the returns to schooling. In terms of the fourth, access to elementary and secondary school quality may be an important determinant of heterogeneous returns to schooling and differences in education choices, and the incorporation of such differences would enhance our understanding of the ways in which various policy changes that alter schooling choices are likely to affect the distribution of achievement, academic attainment, and future earnings. Serious progress along these dimensions is likely to require additional data on the years prior to

kindergarten entry, individual skills and behaviours, the allocation of academic support within schools, and parental time and financial support for learning. The combining of administrative and survey data sets would appear to be a particularly promising way of building a data set with the elements necessary to gain a much better understanding of the distribution of teacher and school effects and the underlying choice frameworks that contribute to the distribution of achievement and future earnings.

References

Aaronson, Daniel, Lisa Barrow, and William Sander (2007) "Teachers and Student Achievement in the Chicago Public High Schools." Journal of Labor Economics 25,no.1 (January):95–135.

Abbring, Jaap H. & Heckman, James J., (2007) "Econometric Evaluation of Social Programs, Part III: Distributional Treatment Effects, Dynamic Treatment Effects, Dynamic Discrete Choice, and General Equilibrium Policy Evaluation," Handbook of Econometrics, in: J.J. Heckman & E.E. Leamer (ed.), Handbook of Econometrics, edition 1, volume 6, chapter 72 Elsevier.

Acemoglu, D and J. Angrist (1999) "How Large are the Social Returns to Education? Evidence from Compulsory Schooling Law", NBER Working Papers 7444,

Adda, Dustmann, Meghir and Robin (2009) "Career Progression and Formal versus On-the-Job Training", IFS Working Paper

Ahn, H., and J. Powell (1993) "Semiparametric Estimation of Censored Selection Models with Nonparametric Selection Mechanism," Journal of Econometrics, 58, 3-29.

Alexander, K.L., D.R. Entwisle, and S.L. Dauber (1996) "Children in motion: School transfers and elementary school performance." Journal of Educational Research 90,no.1 (September/October):3-12.

Altonji, J. and R. Matzkin (2005) "Cross Section and Panel Data Estimators for

Nonseparable Models with Endogenous Regressors" Econometrica, Vol. 73, No. 4 (Jul., 2005), pp. 1053-1102

Altonji, Joseph G., Todd E. Elder and Christopher R. Taber (2005) "Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools," Journal of Political Economy. vol. 113(1): 151-184.

Angrist, Joshua D. and Guido W. Imbens (1995) "Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity" Journal of the American Statistical Association, Vol. 90, No. 430 (Jun., 1995), pp. 431-442

Angrist, Joshua D & Krueger, Alan B, (1991) "Does Compulsory School Attendance Affect Schooling and Earnings?," *The Quarterly Journal of Economics*, MIT Press, vol. 106(4), pages 979-1014, November.

Angrist, Joshua, and Victor Lavy (1999) "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement." Quarterly Journal of Economics 114(2): 533-76.

Arellano, M and S. Bond (1991) "Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations" *The Review of Economic Studies*, 58. pp. 277 – 297.

Attanasio, O. C. Meghir ans A. Santiago (2009) "Education choices in Mexico: using a structural model and a randomized experiment to evaluate PROGRESA.", IFS/ EDEPO Working Paper)

Ballou, D., W. Sanders, and P Wright (2004) "Controlling for Student Background in Value-added Assessment of Teachers." Journal of Educational and Behavioral Statistics v 29 (1): 37-66.

Bayer, Patrick, Fernando Ferreira, and Robert A. McMillan (2007) "A Unified Framework for Measuring Preferences for Schools and Neighborhoods." Journal of Political Economy v. 115 (4): 588-638.

Behrman, J.R., Z. Hrebec, P. Taubman, and T. Wales (1980) "Socioeconomic suc-

cess: A sutdy of the effects of genetic endowments, family environment, and schooling. Amsterdam: North Holland.

Betts, Julian R (1995) "Does School Quality Matter? Evidence from the National Longitudinal Survey of Youth." Review of Economics and Statistics. v 77 (2): 231-250.

Black, Sandra E (1999) "Do Better Schools Matter? Parental Valuation of Elementary Education) " Quarterly Journal of Economics v. 114(2): 577-99.

Blundell, R.W., A. Gosling, H. Ichimura and C. Meghir (2007) "Changes in the Distribution of Male and Female Wages accounting for employment composition using bounds" Econometrica, Vol. 75, No. 2 pp 323–363 (March, 2007),

Blundell, R.W. and J. Powell (2004) "Endogeneity in Semiparametric Binary Response Models" The Review of Economic Studies, Vol. 71, No. 3 (Jul., 2004), pp. 655-67

Blundell, Richard and Frank Windmeijer (1997) "Cluster Effects and Simultaneity in Multilevel Models," Health Economics Letters, vol 6: 439-443.

Bound, John, David A. Jaeger and Regina M. Baker (1995) "Problems with Instrumental Variables Estimation When the Correlation Between theInstruments and the Endogeneous Explanatory Variable is Weak", Journal of the American Statistical Association, Vol. 90, No. 430 (Jun., 1995), pp. 443-450.

Cameron S. V. and J J Hecman (1998) "Life Cycle Schooling and Dynamic Selection Bias: Models and Evidence for Five Cohorts of American Males" The Journal of Political Economy, Vol. 106, No. 2 (Apr., 1998), pp. 262-333

Card, David (1995 "Using Geographic Variation in College Proximity to Estimate the Return to Schooling," in Aspects of Labour Market Behaviour: Essays in Honour of John Vanderkamp, ed. by Louis N. Christofides, E. Kenneth Grant, and Robert Swidinsky. Toronto: University of Toronto Press, 201-222.

Card, David (2001) "Estimating the Return to Schooling: Progress on Some Persistent Econometric Problems", Econometrica, Vol 69.5 pp 1127-1160

Card, David and Alan B. Krueger, (1992) "Does School Quality Matter? Returns to Education and the Characteristics of Public Schools in the United States," Journal of Political Economy, vol. 100(1): pages 1-40.

Carneiro, Pedro, James J. Heckman, Edward Vytlacil (2010) "Evaluating Marginal Policy Changes and the Average Effect of Treatment for Individuals at the Margin", Econometrica Volume 78, Issue 1, Date: January 2010, Pages: 377-394

Chernozhukov, Victor and Han Hong, (2003) "An MCMC Approach to Classical Estimation," Journal of Econometrics

Chesher, A (2003). Identification in Nonseparable Models. *Econometrica*, 71(5), 1405-1441.

Chib, Siddhartha (2001) "Markov chain Monte Carlo methods: computation and inference," Handbook of Econometrics, in: J.J. Heckman & E.E. Leamer (ed.), Handbook of Econometrics, edition 1, volume 5, chapter 57, pages 3569-3649 Elsevier.

Cullen, J.B., B. Jacob, and S. Levitt (2000) "The impact of school choice on student outcomes: An analysis of the Chicago public schools." Working Paper #7888, National Bureau of Economic Research, Cambridge, MA. (September).

Cullen, Julie Barry, Brian A. Jacob, and Steven Levitt (2006) "The Effect of School Choice on Participants: Evidence from Randomized Lotteries," Econometrica, Vol. 74, No. 5: pp. 1191-1230.

Cunha, Flavio, and James J. Heckman (2008) "Symposium on Noncognitive Skills and Their Development: Formulating, Identifying and Estimating the Technology of Cognitive and Noncognitive Skill Formation." Journal of Human Resources, v. 43, iss. 4: 738-82

Cunha, Flavio, James. J. Heckman and Salvador Navarro (2007) "The Identification and Economic Content of Ordered Choice Models with Stochastic Thresholds" International Economic Review Vol. 48, No. 4, November 2007

Cunha, F,., J J Heckman, and S Schennach (2010) "Estimating the Technology of

Cognitive and Noncognitive Skill Formation." NBER Working Paper 15664, forthcoming *Econometrica*.

Darolles, S., Florens, J. -P., & Renault, E (2002). Nonparametric instrumental regression. Working Paper, GREMAQ, University of Social Science, Toulouse

Das, M., W. K. Newey, and Vella (2003) "Nonparametric Estimation of Sample Selection Models," Review of Economic Studies, 70, 33–58.

Dearden, L. C. Emmerson, C. Frayne and C. Meghir (2009) "Conditional Cash Transfers and School Drop out Rates", Journal of Human Resources, Vol 44.4 Fall 2009, pp 827-857

Dearden, Lorraine, Javier Ferri and Costas Meghir (2002) "The Effect Of School Quality On Educational Attainment And Wages." The Review of Economics and Statistics. vol. 84(1): 1-20.

Ding, Weili & Steven Lehrer, 2005) "Class Size and Student Achievement: Experimental Estimates of Who Benefits and Who Loses from Reductions," Working paper.

Eckstein, Z., & Wolpin, K (1989). Dynamic Labour Force Participation of Married Women and Endogenous Work Experience. Review of Economic Studies, 56(3), 375-390.

Epple, D., and R.E. Romano (1998) "Competition between private and public schools, vouchers, and peer-group effects." American Economic Review 88,no.1 (March):33-62.

Fernandez, R., and R. Rogerson (1997) "Education finance reform: A dynamic perspective." Journal of Policy Analysis and Management 16,no.1 (Winter):67-84.

Florens, J. P. J. J. Heckman, C. Meghir, and E. Vytlacil1 (2008) "Identification of treatment effects using control Functions in models with continuous, endogenous Treatment and heterogeneous effects" Econometrica, Vol. 76, No. 5. September, 2008), 1191–1206

Gallipoli, G, C Meghir and G Violante (2008) "Equilibrium Effects of Education

Policies: A Quantitative Evaluation" mimeo University of British Columbia

Griliches, Zvi (1977) "Education, Income, and Ability: Rejoinder," Journal of Political Economy, University of Chicago Press, vol. 85(1), pages 215, February.

Griliches, Zvi (1977) "Estimating the Returns to Schooling: Some Econometric Problems," Econometrica, Econometric Society, vol. 45(1), pages 1-22, January.

Gibbons, Stephen, Stephen Machin, and Olmo Silva (2009) "Valuing School Quality Using Boundary Discontinuities." SERC Discussion Papers 0018, Spatial Economics Research Centre, LSE.

C. Gourieroux, A. Monfort, E. Renault (1993 "Indirect Inference" *Journal of Applied Econometrics*, Vol. 8, Supplement: Special Issue on Econometric Inference Using Simulation Techniques. Dec., 1993), pp. S85-S118

Gronau Reuben (1974 Wage Comparisons–A Selectivity Bias The Journal of Political Economy, Vol. 82, No. 6 (Nov. - Dec., 1974), pp. 1119-1143

Hahn, Jinyong, Petra Todd, and Wilbert Van der Klaauw (2001) "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design." Econometrica. vol. 69(1): 201-09. Hanushek, Eric A., John F. Kain, and Steven G. Rivkin (2004) "Disruption versus Tiebout Improvement: The Costs and Benefits of Switching Schools." Journal of Public Economics, v. 88, iss. 9-10: 1721-46

Hanushek, E.A (1992) "The trade-off between child quantity and quality." Journal of Political Economy 100,no.1 (February):84-117.

Hanushek, E.A., J.F. Kain, J.M. Markman, and S.G. Rivkin (2003) "Does peer ability affect student achievement?" Journal of Applied Econometrics.

Hanushek, E.A., J.F. Kain, and S.G. Rivkin (2001) "Disruption versus Tiebout improvement: The costs and benefits of switching schools." WP 8479, National Bureau of Economic Research (September).

Hanushek, Eric A., John F. Kain, Steve G. Rivkin, and Gregory F. Branch (2007) "Charter School Quality and Parental Decision Making with School Choice." Journal

of Public Economics 91,iss. 5-6:823-48.

Hanushek, Eric A. and Margaret E. Raymond (2005) "Does School Accountability Lead to Improved Student Performance?" Journal of Policy Analysis and Management v. 24 (2): 297-327.

Hanushek, Eric A., Steven G. Rivkin, and Lori L. Taylor, Lori L (1996) "Aggregation and the Estimated Effects of School Resources." The Review of Economics and Statistics. vol. 78(4): 611-27.

Harmon, C., & Walker, I (1995). Estimates of the Economic Return to Schooling for the United Kingdom. American Economic Review, 85(5), 1278-1286.

Harris, Douglas N. and Tim R. Sass (2009) "The Effects of NBPTS-Certified Teachers on Student Achievement." Journal of Policy Analysis and Management. v. 28, iss. 1: 55-80.

Heckman James (1974 Shadow Prices, Market Wages, and Labor Supply Econometrica, Vol. 42, No. 4 (Jul., 1974), pp. 679-694

Heckman, J.J (1979) "Sample Selection Bias as a Specification Error", Econometrica 47, 153–162. [326,336]

Heckman, J.J (1990) "Varieties of Selection Bias," The American Economic Review, Papers and Proceedings of the Hundred and Second Annual Meeting of the American Economic Association 80, pages 313-318.

Heckman, J.J. ., and B. Honore (1990) "The Empirical Content of the Roy Model," Econometrica, 58, 1121–1149.

Heckman, J.J. , R. LaLonde, and J. Smith (1999) "The Economics and Econometrics of Active Labour Market Programmes," in Handbook of Labour Economics, Vol. 3a, ed. by O. Ashenfel- ter and D. Card. Amsterdam: North-Holland, 1865–2097

Heckman, James, Anne Layne-Farrar, and Petra Todd (1996) "Human Capital Pricing Equations with an Application to Estimating the Effect of Schooling Quality on Earnings." The Review of Economics and Statistics. vol. 78(4): 562-610.

Heckman, J. J. , L. Lochner and C. Taber (1998 "Explaining Rising Wage Inequality: Explanations With A Dynamic General Equilibrium Model of Labor Earnings With Heterogeneous Agents," Review of Economic Dynamics, Elsevier for the Society for Economic Dynamics, vol. 1(1), pages 1-58, January

Heckman, J. J. , L. Lochner and P. Todd (2003) "Fifty Years of Mincer regressions" IZA Discussion Paper No. 775

Heckman, J.J. and S. Navarro (2007) "Dynamic discrete choice and dynamic treatment effects" Journal of Econometrics 136 341–396

Heckman, J. and R. Robb (1985), Alternative methods for evaluating the impact of interventions. in: J. Heckman and B. Singer, eds. Longitudinal analysis of labor market data, Econometric Society monograph series. Cambridge University Press, New York).

Heckman, J., & Sedlacek, G (1985). Heterogeneity, Aggregation, and Market Wage Functions: An Empirical Model of Self-selection in the Labor Market. *Journal of Political Economy*, 93(6), 1077-1125.

Heckman, J.J and B. Singer (1984) "A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data" Econometrica Vol. 52, No. 2 (Mar., 1984), pp. 271-320

Heckman, J.J., and J.A. Smith (1999) "The pre-programme earnings dip and the determinants of participation in a social programme: Implications for simple programme evaluation strategies." The Economic Journal 109(July):313-348.

Heckman, J. J., S. Urzua and E. Vytlacil (2006a) "Understanding Instrumental Variables in Models with Essential Heterogeneity," Review of Economics and Statistics, 88(3)

Heckman, J. J., S. Urzua and E. Vytlacil (2006b) "Instrumental Variables in Models with Multiple Outcomes: The General Unordered Case", mimeo University of Chicago.

Heckman, James J. and Edward Vytlacil (2005) Structural Equations, Treatment

Effects, and Econometric Policy Evaluation, Econometrica Volume 73, Issue 3, Date: May 2005, Pages: 669-738.

Hsiao, Cheng (2003) “Analysis of Panel Data 2nd Edition. Cambridge University Press.

Imbens, G. W. and J. D. Angrist (1994) "Identification and Estimation of Local Average Treatment Effects," Econometrica, 62, 467-475.

Imbens, G. and W. Newey (2009) "Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity” Econometrica Volume 77, Issue 5, September 2009, Pages: 1481-1512

Ingersoll, G.M., J.P. Scamman, and W.D. Eckerling (1989) "Geographic mobility and student achievement in an urban setting." Educational Evaluation and Policy Analysis 11,no.2 (Summer):143-149.

Jepsen, Christopher and Steven Rivkin (2009) "Class Size Reduction and Student Achievement: The Potential Tradeoff between Teacher Quality and Class Size." Journal of Human Resources vol. 44(1): 223-250.

Kain, J.F., and D.M. O’Brien (1998) "A longitudinal assessment of reading achievement: Evidence for the Harvard/UTD Texas Schools Project." UTD Texas Schools Project, University of Texas at Dallas (April 1998).

Kane, Thomas J., and Douglas O. Staiger (2008) "Are Teacher-Level Value-Added Estimates Biased? An Experimental Validation of Non-Experimental Estimates." Harvard University, (mimeo March).

Katz, L. and K. Murphy (1992 , "Changes in relative wages, 1963-1987: Supply and demand factors", Quarterly Journal of Economics 107 1992 , 35-78.

Keane Michael P. and Kenneth I. Wolpin (1997 The Career Decisions of Young Men The Journal of Political Economy Vol. 105, No. 3 (June 1997), pp. 473-522.

Kerbow, D (1996) "Patterns of urban student mobility and local school reform." Journal of Education for Students Placed at Risk 1,no.2:147-169.

Kennan, John and James R. Walker (2009) "The Effect of Expected Income on Individual Migration Decisions." University of Wisconsin working paper.

Kitagawa, Toru (2010)“Testing for Instrument Independence in the Selection Model”, CEMMAP working paper, February 2010.

Konstantopoulos, Spyros (2008) “Do Small Classes Reduce the Achievement Gap Between Low and High Achievers, Evidence from Project STAR?” Elementary School Journal, 108: 275-291.

Krueger, Alan B (1999) "Experimental Estimates Of Education Production Functions." The Quarterly Journal of Economics. vol. 114(2): 497-532,

Lazear, Edward (2001) "Educational Production.” Quarterly Journal of Economics. 116: 3: 777-803.

Lee, Donghoon (2005) "An Estimable Dynamic General Equilibrium Model of Work, Schooling and Occupational Choice”, International Economic Review, 46, February, 2005, 1-34.

Lee, D. and K. Wolpin (2006) "Intersectoral Labor Mobility and the Growth of the Service Sector”, Econometrica, 47, January, 1-46.

Lerman, S., and C. Manski (1981) "On the Use of Simulated Frequencies to Approximate Choice Probabilities," in Structural Analysis of Discrete Data with Econometric Applications, ed. by C. Manski and D. McFadden. Cambridge: MIT Press, 305-319.

Mare, R.D (1980) "Social background and school continuation decisions." Journal of the American Statistical Association 75,no.370 (June):295-305.

Mayer, S.E (1997) What money can’t buy: Family income and children’s life chances. Cambridge, MA: Harvard University Press.

Magnac, Thierry, and David Thesmar (2002), “Identifying Dynamic Discrete Decision Processes”, Econometrica, Vol. 70, No. 2 (March, 2002), 801–816

Manski, C (1994) "The Selection Problem,” in Advances in Econometrics, Sixth

World Congress, Vol 1, ed. by C. Sims. Cambridge, U.K.: Cambridge University Press, 143–170. [327,329,331]

Manski, C. and J. Pepper (2000) "Monotone Instrumental Variables: With Application to the Returns to Schooling," Econometrica, 68, 997–1010. [327,332]

McFadden, Daniel (1989) "Method of Simulated Moments for Estimation of Discrete Response Models Without Numerical Integration" Econometrica, Vol. 57, No. 5 (Sep., 1989), pp. 995-1026

Meghir, Costas, and Mårten Palme (2005 "Educational Reform, Ability, and Family Background." American Economic Review, 95(1): 414–424

Meghir, C (2006) "Dynamic Models for Policy Evaluation", Advances in economics and econometrics: theory and applications, 9th World Congress of the Econometric Society, UCL, London, Richard Blundell, Whitney K. Newey, Torsten Persson eds.

Mincer, J (1958) "Investment in Human Capital and Personal Income Distribution," Journal of Political Economy, 66(4):281-302.

Mincer, J (1974) Schooling, Experience, and Earnings, New York: NBER Press.

McKee, Graham, Steven Rivkin, and Katharine R. E. Sims (2010) "Disruption, Achievement and the Heterogeneous Benefits of Smaller Classes " NBER working paper number.

Morris, C (1983) "Parametric Empirical Bayes Inference: Theory and Applications (with discussion)." Journal of the American Statistical Association v 78: 47-65.

Neal, Derek (1997) "The Effects of Catholic Secondary Schooling on Educational Achievement." Journal of Labor Economics, Part 1 v. 15(1): 98-123.

Nechyba, T.J (2000) "Mobility, targeting, and private-school vouchers." American Economic Review 90,no.1 (March):130-146.

Nerlove, Marc (1971) "Further evidence on the estimation of dynamic economic relations from a time series of cross sections" *Econometrica*, 39(2): 359-382.

Nickell, S (1981) "Biases in dynamic models with Fixed effects" *Econometrica*,

49(6): 1417-1426.

Newey, W. and J. Powell (2003) "Instrumental Variable Estimation of Nonparametric Models" Econometrica, Vol. 71, No. 5 (Sep., 2003), pp. 1565-1578.

Newey, W. J. Powell and F. Vella (1999) "Nonparametric Estimation of Triangular Simultaneous Equations Models" Econometrica, Vol. 67, No. 3 (May, 1999), pp. 565-603.

Orcutt, G. H., and A. G. Orcutt (1968 Incentive and Disincentive Experimentation for Income Maintenance Policy Purposes," American Economic Review, 58, 754-772.

Pakes, A., and Pollard R (1989 "The Asymptotic Distribution of Simulation Experiments," Econometrica, 57, 1027-1057.

Pong, S. and A.M. Pallas (2001) "Class Size and Eighth-Grade Math Achievement in the United States and Abroad." Educational Evaluation & Policy Analysis 23(3):251-273.

Raudenbush, S.W., & Bryk, A.S (2002) Hierarchical linear models: Applications and data analysis methods (2nd ed.). Thousand Oaks, CA: Sage.

Rivkin, S.G., E.A. Hanushek, and J.F. Kain (2001) "Teachers, schools, and academic achievement." Working Paper No. 6691, National Bureau of Economic Research (revised)

Rivkin, Steven G., Eric A. Hanushek, and John F. Kain (2005) "Teachers, schools, and academic achievement." Econometrica 73,no.2 (March):417-458.

Robert, C P and G. Casella (1999) "Monte Carlo Statistical Methods", Springer Verlag - New York.

Rockoff, Jonah E (2004) "The impact of individual teachers on student achievement: Evidence from panel data." American Economic Review 94,no.2 (May): 247-252.

Rothstein, Jesse (2009) "Student sorting and bias in value added estimation: Selection on observables and unobservables," NBER Working Paper number 14666.

Roy A D (1951) "Some Thoughts on the Distribution of Earnings", Oxford Economic Papers, New Series, Vol. 3, No. 2 (Jun., 1951), pp. 135-146

Rust, John (1987) "Optimal Replacement of GMC Bus Engines: An Empirical Model of Harold Zurcher", Econometrica, 55 999-1033

Rust, J. (1994) "Structural Estimation of Markov Decision Processes," in Handbook of Econometrics, Vol. 4, ed. by R. Engle and D. McFadden. Amsterdam: North Holland, pp. 3081–3143.

Schacter, J (2001a) "Geographic mobility: March 1999 to March 2000." In Current Population Reports. Washington, DC: U.S. Census Bureau

Schacter, J (2001b) "Why people move: Exploring the March 2000 Current Population Series." In Current Population Reports. Washington, DC: U.S. Census Bureau

Staiger Douglas and James H. Stock (1997) "Instrumental Variables Regression with Weak Instrument"s", Econometrica, Vol. 65, No. 3 (May, 1997), pp. 557-586.

Stinchcombe , A.L (1969) "Environment: The Cumulation of Effects." Harvard Educational Review 39,no.3 (Summer):511-522.

Straszheim, M (1987) "The Theory of Urban Residential Location", in Handbook of Regional and Urban Economics, edited by E. S. Mills, Amsterdam: North-Holland: 717-757.

Su, Che-lin and Kenneth Judd (2008) "Constrained Optimisation Approaches to Estimation of Structural Models", mimeo Stanford

Theil, Henri (1954)." Linear Aggregation of Economic Relations". North Holland Publishing Co.

Todd Petra E. and Kenneth I. Wolpin (2003) "On the Specification and Estimation of the Production Function for Cognitive Achievement. Economic Journal: F3-F33.

Vytlacil, E. (2002) "Independence, Monotonicity, and Latent Index Models: An Equivalence Result", Econometrica,Vol. 70, No. 1 (Jan., 2002), pp. 331-341.

Vytlacil, E. J. (2006) "Ordered Discrete-Choice Selection Models and Local Aver-

age Treatment Effect Assumptions: Equivalence, Nonequivalence, and Representation Results," Review of Economics and Statistics 88 (2006), 578-81.

Weimer, D.L., and M.J. Wolkoff (2001) "School performance and housing values: Using non-contiguous district and incorporation boundaries to identify school effects." National Tax Journal 54,no.2 (June):231-253.

Wildasin, D. E (1987) "Theoretical Analysis of Local Public Economics." in Handbook of Regional and Urban Economics, edited by E.S. Mills. Amsterdam: North-Holland:1131-1178.

Willis R. and S. Rosen (1979) "Education and Self Selection", Journal of Political Economy, 87, pp S7-S36.

Wolpin, K. I (1992) "The Determinants of Black-White Differences in Early Employment Careers: Search, Layoffs, Quits and Endogenous Wage Growth," Journal of Political Economy, 100(3), 835-560.