JOURNAL OF APPLIED ECONOMETRICS
J. Appl. Econ. **19**: 827–849 (2004)
Published online 11 March 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/jae.744

LEARNING ABOUT HETEROGENEITY IN RETURNS TO SCHOOLING

GARY KOOP[a]* AND JUSTIN L. TOBIAS[b]
[a] *Department of Economics, University of Leicester, UK*
[b] *Department of Economics, University of California, Irvine, USA*

SUMMARY

Using data from the National Longitudinal Survey of Youth (NLSY) we introduce and estimate various Bayesian hierarchical models that investigate the nature of unobserved heterogeneity in returns to schooling. We consider a variety of possible forms for the heterogeneity, some motivated by previous theoretical and empirical work and some new ones, and let the data decide among the competing specifications. Empirical results indicate that heterogeneity is present in returns to education. Furthermore, we find strong evidence that the heterogeneity follows a continuous rather than a discrete distribution, and that bivariate normality provides a very reasonable description of individual-level heterogeneity in intercepts and returns to schooling.

1. INTRODUCTION

Recent research in labour economics has investigated the existence of individual-level heterogeneity in returns to education and has discussed the use and interpretation of instrumental variable (IV) methods when schooling returns are heterogeneous. In thorough reviews of this literature, Card (1999, 2001) observes that most of the recent empirical work in this area has exploited supply-side instrumental variables, and also notes that the IV point estimates in these studies tend to exceed their OLS counterparts. While several explanations exist to rationalize this result, Card offers a new explanation. Given the interpretation of IV estimates when returns to 'treatment' are heterogeneous (e.g. Imbens and Angrist, 1994), the IV estimates reported in previous work may simply recover the returns to schooling for particular subgroups of the population that 'comply' with the instruments.[1] Since different instruments affect the schooling decisions of different subpopulations, and these subpopulations can have different returns to education, it is not surprising that IV estimates differ across studies. In addition, if marginal returns to schooling are higher on average for those individuals 'complying' with the instruments, then IV point estimates should also exceed their OLS counterparts.

The above explanation reveals that the IV point estimate only provides us with a specific (and potentially undesired) feature of the heterogeneity distribution. In this paper we seek to escape this limitation by investigating heterogeneity in returns to education in a more direct way. Our approach aims to characterize the distribution of heterogeneity in the population without needing

* Correspondence to: Gary Koop, Department of Economics, University of Leicester, University Road, Leicester, LE1 7RH, UK. E-mail: gary.koop@leicester.ac.uk
[1] For example, if the instrument employed is an indicator denoting if a college is present in an individual's county, then IV will estimate the return to schooling for those attending college provided a college is in the county, but otherwise would not. See Kling (2001) for related discussion.

Received 25 September 2002
Revised 1 May 2003

to rely on instrumental variables or natural experiments. In particular, we exploit time-variation in educational attainment to test for the existence of and model the distribution of heterogeneity in returns to education.

The models we use involve two stages. In the first stage, a linear specification relates log wages to years of schooling (and other explanatory variables). Heterogeneity is incorporated by allowing the intercept and slope of this linear relationship to vary across individuals. In the second stage, a probability distribution is used to model this variation. Our models differ in the specification of this heterogeneity distribution. We consider models with no heterogeneity (i.e. the second stage distribution is degenerate), normally distributed heterogeneity (i.e. the intercept and return to schooling are normally distributed in the population) or discretely distributed heterogeneity (i.e. there exist G distinct groups of individuals and within each group returns to schooling are identical). All of these competing forms have some precedent in the literature. We also estimate other more general specifications for the heterogeneity distribution, and introduce explanatory variables into the second stage of our hierarchical model to see if observable characteristics might help to explain the unobserved heterogeneity across individuals.

While the methods employed in this paper enable us to learn about the form of the heterogeneity distribution directly, our identification strategy necessarily requires longitudinal variation in educational attainment for some individuals in our sample. While there is some precedent for exploiting this type of variation (e.g. Angrist and Newey, 1991), it is important to recognize that most empirical work in this area restricts attention to the earnings history of older workers who have completed their education. Thus, our approach takes advantage of the availability of wage observations for individuals whose transition from school to work is not 'smooth', while most work in this area simply conditions on those observations where schooling is time-invariant. Though the models we propose are somewhat non-standard in light of the previous literature, our view is that the benefits afforded by estimating models that exploit time-variation in educational attainment are great enough to merit a careful empirical investigation. In particular, we will be able to formally *test for the existence of* and *learn about the nature of* heterogeneity in returns to education. As shown in the previous section, the results we obtain are in line with those reported in previous work, and the extent of heterogeneity we find is both plausible and arguably reconciles the disparity of IV estimates found in earlier studies.

We take panel data from the National Longitudinal Survey of Youth (NLSY) to estimate models that permit individual-specific intercept and return to education parameters, following Card's (2001) observation that such a specification emerges from a theoretical model accounting for forces of both supply and demand. Although the length of our panel is rather short, this does not create a significant problem for us as we employ a Bayesian approach which provides exact finite sample results. Furthermore, estimates of the individual-level parameters obtained from our hierarchical model incorporate not only information from the outcomes of that individual, but also incorporate information obtained from the other individuals in the sample. In this sense, the final individual-level parameter estimates are not solely determined by data from the individual, but instead balance longitudinal information from the outcomes of the given individual and cross-section information obtained from the parameter estimates of all individuals.

The outline of the paper is as follows. In the following section, we present a general hierarchical model for investigating heterogeneity in returns to schooling. In Section 3 we focus on some special cases of this general framework that have appeared in the literature, and discuss the interpretation of each model. Section 4 describes the data and the empirical results are reported in Section 5. The

paper concludes in Section 6. Technical details, including development of Markov Chain Monte Carlo (MCMC) algorithms and further discussion of the priors are included in appendices.

2. LEARNING ABOUT THE NATURE OF UNOBSERVED HETEROGENEITY

To focus the discussion, it is useful to present at this stage the general form for the class of models that we use in this paper. Precise details are given in the next section. Our models allow for individual-level variation in the intercept and return to education:[2]

$$y_{it} = \alpha_i + \beta_i s_{it} + \gamma z_{it} + \varepsilon_{it} \tag{1}$$

$$\alpha_i, \beta_i | \lambda, w_i \overset{ind}{\sim} f(\alpha_i, \beta_i | \lambda, w_i) \tag{2}$$

where y_{it} denotes the log hourly wage of individual i at time (year) t, s_{it} denotes years of schooling completed, α_i and β_i are individual-specific intercept and return to schooling parameters (respectively), and z_{it} represents a set of time-varying characteristics (e.g. labour market experience, etc.). We assume, conditional on the person-specific effects and adequate controls for time effects (in z), that ε is an i.i.d. normal disturbance term.[3] We also assume, as is often done in this literature following early work by Mincer (1974), Becker and Chiswick (1966) and Heckman and Polachek (1974) (among many others), that the log-wage is linear in schooling.[4]

Equation (2) introduces the heterogeneity distribution and, at this stage, we represent this generally by some bivariate distribution f. The vector λ contains the parameters of this distribution, and w_i are time-invariant variables such as measured cognitive ability and family characteristics that might play a role in explaining heterogeneity across individuals. The primary goal of this paper is to learn about the nature of the underlying heterogeneity across individuals, and thus to learn about f and the second-stage parameters λ. We do this by assuming different forms of heterogeneity through different specifications for f, and tie these competing forms to previous specifications or assumptions employed in the past literature. Our goal is to determine if such heterogeneity is present across individuals, and if so, how best to model it.

2.1. Previous Issues Raised in the Literature

The model given in (1), through its ability to allow for heterogeneity in intercepts and returns to education, has been interpreted as a model that allows marginal costs of and marginal returns to education to vary across individuals in the population (see, e.g., Card, 2001). Several issues arise in the analysis of such a model, and these issues have received substantial attention in the literature. Most importantly, issues of the endogeneity of schooling choice as well as measurement

[2] Some studies have investigated the issue of time-varying returns to schooling over a similar period of study. We do not, as yet, investigate this issue, but defer it as the subject of future work.

[3] Note that this assumption could be relaxed by adding mixing variables to the error variance to obtain, for example, Student-t errors (Carlin and Polson, 1991; Geweke, 1993).

[4] Numerous studies have investigated the existence of 'jumps' in the schooling–log wage relationship upon degree completion, or 'sheepskin' effects (e.g. Hungerford and Solon, 1987; Belman and Heywood, 1991; Heywood, 1994; Heckman *et al*. 1996; Jaeger and Page, 1996). In this paper, we restrict our attention to the linear-in-schooling model and focus on heterogeneity in returns in the context of this widely-used specification.

error in schooling[5] have been investigated in recent work. If these problems exist then biased and inconsistent estimates of underlying structural parameters of interest could result.

We first note that the representation in (1) might be regarded as a model that captures heterogeneity either in 'structural form' or 'reduced form'. If it is structural, the interpretation of results is straightforward and sensible. But, even if our model is interpreted as a reduced form one, our view is that direct analysis of this relationship is of interest for a variety of reasons. First, we note that since many studies in the empirical education literature have directly analysed the reduced form relationship in (1), it is useful to begin our investigation here for the sake of comparison with the existing literature, and simply to offer a starting point and a new way to look at issues of heterogeneity in returns to education. Second, and perhaps most importantly, we argue that the use of panel data helps to mitigate concerns regarding endogeneity of schooling.[6] Unobservable factors like ability and motivation whose presence is typically argued to violate the mean-independence assumption $E(\varepsilon_{it}|s_{it}, z_{it}) = 0$ are directly controlled for through the individual effects α_i in (1), given that these factors do not vary with time. As such, it is seemingly reasonable to maintain $E(\varepsilon_{it}|s_{it}, z_{it}) = 0$ as well as assume independence across observations conditioned on these individual factors. If these assumptions are correct, then endogeneity problems will not plague our empirical results.

Finally, though perhaps less importantly, it may prove to be useful to look a bit deeper at the relationship between the reduced form and structural parameters. One such structural model, which permits endogeneity of schooling choice,[7] would be written as follows:

$$\begin{aligned} y_{it} &= \tilde{\alpha}_i + \tilde{\beta}_i s_{it} + \tilde{\gamma} z_{it} + \varepsilon_{it} \\ s_{it} &= \tilde{\delta}_i + \tilde{\eta} v_{it} + u_{it} \end{aligned} \tag{3}$$

In the above, $\tilde{\beta}_i$ is the *structural* return to schooling parameter, $\tilde{\delta}_i$ is an individual-specific intercept that may capture variation in costs of and tastes for schooling across individuals and v_{it} a vector of explanatory variables affecting schooling quantity. We assume that the distribution of $\tilde{\delta}_i$ is independent across individuals and also independent of the explanatory variables z and v. If ε_{it} and u_{it} are independent, then estimates obtained from (1) and (2) can be interpreted as structural parameters. However it is still possible (though perhaps unlikely after controlling for $\tilde{\alpha}_i$) that ε_{it} and u_{it} are sufficiently correlated to make endogeneity concerns important.

Assuming (3) is the structural model, (1) can be interpreted as the *reduced form* model (i.e. it is the *conditional distribution of y given s and other exogenous variables*). The relationship between the parameters of the structural and reduced form models can be obtained by working out the

[5] We do not investigate the issue of measurement error and assume that education is measured correctly in our final sample. As described in Section 4, we are careful to exclude individuals whose education is clearly misreported or obviously suspect, which undoubtedly helps to mitigate the problem. Using NLSY data, Blackburn and Neumark (1995, p. 228) summarize their findings regarding the severity of measurement error and state: '... once test scores are included in the regression, specifications tests find little evidence that schooling is either endogenous or measured with error, or that ability is measured with error by the test scores.' Since our models control for person-specific effects and test scores, these issues may not be problematic in our analysis.

[6] The assumption of ignorable endogeneity conditioned on adequate controls has been implicitly made numerous times in the progression of this literature. See, for example, Ashenfelter and Mooney (1968), Hansen *et al*. (1970), Hause (1972), Hungerford and Solon (1987), Lam and Schoeni (1993), Blackburn and Neumark (1995), Cawley *et al*. (1997) and Heckman and Vytlacil (2001), who do not instrument for schooling given a rich set of control variables.

[7] Such models which permit endogeneity of schooling have appeared numerous times in the literature—see, for example, Angrist and Krueger (1991, p. 997, eqs (1) and (2)).

conditional distribution of y given s (and other explanatory variables) implied by (3). If we make the common assumption that

$$\begin{bmatrix} \varepsilon_{it} \\ u_{it} \end{bmatrix} \sim N(0, \Sigma), \quad \text{where } \Sigma = \begin{bmatrix} \sigma_\varepsilon^2 & \sigma_{\varepsilon u} \\ \sigma_{\varepsilon u} & \sigma_u^2 \end{bmatrix}$$

then we can use (3) to show that

$$\begin{bmatrix} y_{it} \\ s_{it} \end{bmatrix} \Big| \phi, z_{it}, v_{it} \sim N(\Gamma^{-1}\mu, \Gamma^{-1}\Sigma[\Gamma^{-1}]')$$

where Γ^{-1} is a 2×2 upper triangular matrix with diagonal elements equal to one and (1,2) element $\tilde{\beta}_i$, ϕ denotes all parameters in our model and

$$\Gamma^{-1}\mu = \begin{bmatrix} \tilde{\alpha}_i + \tilde{\beta}_i\tilde{\delta}_i + \tilde{\gamma}z_{it} + \tilde{\beta}_i\tilde{\eta}v_{it} \\ \tilde{\delta}_i + \tilde{\eta}v_{it} \end{bmatrix}, \quad \Gamma^{-1}\Sigma[\Gamma^{-1}]' = \begin{bmatrix} \sigma_\varepsilon^2 + 2\tilde{\beta}_i\sigma_{\varepsilon u} + \tilde{\beta}_i^2\sigma_u^2 & \sigma_{\varepsilon u} + \tilde{\beta}_i\sigma_u^2 \\ \sigma_{\varepsilon u} + \tilde{\beta}_i\sigma_u^2 & \sigma_u^2 \end{bmatrix}$$

Working out the conditional of y given s (and other explanatory variables) obtained from this bivariate normal, we find

$$y_{it}|s_{it}, z_{it}, v_{it}, \phi \sim N\left[\tilde{\alpha}_i + \tilde{\beta}_i s_{it} + \tilde{\gamma}z_{it} - \frac{\sigma_{\varepsilon u}}{\sigma_u^2}(\tilde{\delta}_i + \tilde{\eta}v_{it} - s_{it}), \sigma_\varepsilon^2(1 - \rho_{\varepsilon u}^2)\right]$$

By examining the coefficients on s_{it}, we can work out the relationship between the reduced form and structural form parameters. Most importantly we obtain:

$$\beta_i = \tilde{\beta}_i + \frac{\sigma_{\varepsilon u}}{\sigma_u^2}, \quad \alpha_i = \tilde{\alpha}_i - \left(\frac{\sigma_{\varepsilon u}}{\sigma_u^2}\right)\tilde{\delta}_i$$

Hence, provided the instruments v_{it} are included in z_{it} in (1),[8] *the variability in reduced-form slope coefficients across individuals equals the variability of the underlying structural coefficients across individuals*. Statistically speaking, the mean of our heterogeneity distribution obtained from analysis of the conditional distribution $y|s, z, v$ may be wrong, but all of its other characteristics (i.e. its variance, its shape, etc.) remain correct. Thus, we can still test for and quantify the extent of heterogeneity in slope coefficients across individuals by working with the conditional distribution of log wages. In other words, if our primary goal is to learn about the existence and extent of variation in returns to education across individuals, it is enough to analyse the conditional distribution $y_{it}|s_{it}, z_{it}, v_{it}, \phi$, rather than the joint distribution given in (3). This motivates the class of models whose general form is given in (1) as our primary focus.

[8] In some of our models we include characteristics like number of siblings and indicators for residence in a broken home at age 14. These variables are correlated with the quantity of schooling attained but arguably have no structural effect on wages conditioned on the other controls. Hence, we might potentially interpret these models as containing instruments in the regression. Relatedly, Blackburn and Neumark (1992) also used family characteristics as instruments in wage equations, arguing these characteristics have no direct structural dependence with wages, but only affect wages indirectly through education and ability. Griliches (1979) reported a similar result. Even without this result, we maintain that the availability and use of panel data helps to mitigate these endogeneity concerns.

3. A SET OF COMPETING MODELS

A complete description of the class of models we consider is provided in this section. To this end, we write (1) more compactly and expand it to include the prior:[9]

$$y_{it}|x_{it}, z_{it}, \theta_i, \gamma, \sigma_\epsilon^2 \overset{ind}{\sim} N(x_{it}\theta_i + z_{it}\gamma, \sigma_\epsilon^2), \quad i = 1, 2, \ldots, n, \ t = 1, 2, \ldots, T_i \tag{4}$$

$$\theta_i|\lambda, w_i \overset{ind}{\sim} f(\theta_i|\lambda, w_i) \tag{5}$$

$$\gamma|\underline{\mu}_\gamma, \underline{V}_\gamma \sim N(\underline{\mu}_\gamma, \underline{V}_\gamma) \tag{6}$$

$$\lambda|\underline{\lambda} \sim g(\underline{\lambda}) \tag{7}$$

$$\sigma_\epsilon^{-2}|\underline{s}_\epsilon^{-2}, \underline{\nu}_\epsilon \sim G(\underline{s}_\varepsilon^{-2}, \underline{\nu}_\varepsilon) \tag{8}$$

where we use the notation $\underline{\cdot}$ to denote terminal hyperparameters assigned by the researcher, and $f(\theta_i|\lambda, w_i)$ is a hierarchical prior that depends on a parameter vector λ and a $1 \times k_w$ row vector w_i. In some cases, w_i will simply include a constant term. In (4) we have also defined

$$\theta_i = \begin{bmatrix} \alpha_i \\ \beta_i \end{bmatrix}, \quad x_{it} = [1 \quad s_{it}]$$

As stated previously our primary goal is to learn about f and λ. To that end, we consider a variety of alternatives for f—most of these guided by previous research in this area—and let the data decide among the competing specifications. In particular, we focus primarily on five competing models, and enumerate these in Table I.

Model 1 represents our most restrictive specification and imposes that all individuals share a common intercept and marginal return to schooling. In statistical language this model specifies the heterogeneity distribution, f, as being degenerate at the point (α, β).

In **Model 2** we allow slopes and intercepts to differ across individuals, but impose a degree of similarity across individuals by assuming that effects are drawn from the same normal population. Such hierarchical or 'random effects' models have been suggested by or employed in past work in the schooling literature by Becker and Chiswick (1966) and Chiswick (1974), among others. More recently, theoretical issues in an elaborated version of this model were described in Heckman and Vytlacil (1998).

In practice we will also consider generalizing Model 2 by incorporating time-invariant characteristics (e.g. ability and family background variables) into the second stage of the hierarchy.[10]

Table I. Alternate heterogeneity models

	Model 1	Model 2	Model 3	Model 4	Model 5
Restriction/ assumption	$\alpha_i = \alpha$ $\beta_i = \beta$	$\theta_i \sim N(\theta_0, \Sigma)$	$\alpha_i \sim N(\alpha, \sigma_\alpha^2)$ $\beta_i = \beta$	$\theta_i = \theta_g^0$ with probability π_g	$p(\theta_i) = \sum_{g=1}^G \pi_g \phi(\theta_i; \theta_g^0, \Sigma_g)$

[9] In our notation, $N(a, b)$ denotes a normal distribution with mean a and variance b, and $G(s_\epsilon^{-2}, \nu_\epsilon)$ denotes a Gamma density, parameterized as in Poirier (1995, p. 100).

[10] Several studies have addressed the issue of returns to schooling varying with observable characteristics—particularly measures of cognitive ability. See, for example, Hause (1972), Blackburn and Neumark (1993), Murnane *et al*. (1995), Grogger and Eide (1995), Heckman and Vytlacil (2001) and DiNardo and Tobias (2001).

When these characteristics are included we denote the resulting model as **Model 2W**. Letting

$$w_i = \begin{bmatrix} w_{\alpha i} & 0 \\ 0 & w_{\beta i} \end{bmatrix}, \quad \theta_0 = \begin{bmatrix} \theta_\alpha \\ \theta_\beta \end{bmatrix}$$

we define Model 2W as the model imposing $\theta_i | w_i, \theta_0, \Sigma \overset{ind}{\sim} N(w_i\theta_0, \Sigma)$.

We also note that a formal comparison of Model 2 against Model 1 does not necessarily test for the existence of heterogeneity in returns to schooling, simply because Model 2 also permits individual-specific intercepts. In other words, a preference for Model 2 over Model 1 simply indicates a need to control for *some form of heterogeneity at the individual level*. To enable us to perform the desired test, we introduce **Model 3** that permits heterogeneity in intercepts only. By comparing Model 3 against Model 2, we are able to test if the data favour a model that *additionally* permits heterogeneity in returns to schooling to one that only permits heterogeneity through baseline differences intercepts.

While Models 2 and 3 clearly generalize the common parameter assumption in Model 1, a potential limitation of these second-stage specifications is their reliance on normality. That is, the true form of the heterogeneity may be something quite different than a normal distribution, leading us to potentially mischaracterize the nature of the heterogeneity.

One alternative, which was suggested in earlier work by Heckman and Singer (1984), is to place a discrete distribution on θ_i. As pointed out in their paper, and also described in Allenby and Rossi (1999), it is possible that the distribution of the heterogeneity can be approximated quite accurately by a discrete distribution with a suitable number of support points. In **Model 4**, we discretize the support of θ_i and thus approximate the unknown heterogeneity distribution with a discrete distribution. The parameters of interest include values of the support points θ_g^0, their associated probabilities π_g, as well as the appropriate *number of points* G. In the latter sense, Model 4 actually denotes many different models, where these models are indexed by different values for G.[11]

It is worthwhile to note that the assumption of discrete heterogeneity has some precedent in this literature. For example, the approach of Ichino and Winter-Ebmer (1999) could be rationalized as one that supposes the existence of four different 'types' of individuals in the population, where each individual has a low or high marginal cost of and marginal return to education.

Model 5 is our most flexible representation of the distribution of θ_i, and in this model we assume that the heterogeneity distribution follows a finite mixture of normals. This offers a flexible modelling alternative that is capable of capturing a variety of possible forms for the heterogeneity. Like Model 4, Model 5 assumes that the population is comprised of a discrete number of different groups, but unlike Model 4, it permits individuals within each group (or component of the mixture) to possess different intercepts and returns to education.

4. THE DATA

We obtain data to fit our competing models from the National Longitudinal Survey of Youth (NLSY). The NLSY is a rich panel study of 12,686 individuals in the USA ranging in age from

[11] We also describe and estimate a **Model 4W** that introduces time-invariant characteristics w_i to the second stage. This model amounts to a (restricted) finite mixture model for log wages coupled with an ordered probit model to explain the component assignment mechanism. Details can be obtained from the complete technical appendix: http://orion.uci.edu/~jtobias/ktappend.pdf.

14–22 as of the first interview date in 1979. Importantly for our purposes, the NLSY contains detailed information on the wages, educational attainment, family characteristics and test scores of the sampled individuals. For this application we use a version of the NLSY that allows us to obtain an earnings history until 1993.

To abstract from selection issues in employment and to remain consistent with the majority of the literature, we focus our attention on the outcomes of white males in the NLSY. At the early years of the sample, many of these individuals are still enrolled in school, thus potentially calling into question the use of wage data for a very young set of workers. To ensure that the hourly wage variable does a reasonably good job at picking up the earnings *potential* of the sampled individuals, we restrict attention to those individuals who are active in the labour force for a good portion of each year. Specifically, we restrict the sample to white males who are at least 16 years of age in the given year, who reported working at least 30 weeks a year and at least 800 hours per year. We also delete observations when the reported hourly wage is less than \$1 or greater than \$100 per hour, when education decreases across time for an individual, or when the reported change in years of schooling over time is not consistent with the change in time from consecutive interviews. Thus, we are careful to delete individuals whose education or wage is clearly misreported.

Given this sample selection scheme, we obtain data on $N = 2178$ individuals for a total (denoted NT) of $NT = 17{,}919$ person-year observations. Our time-varying characteristics include potential labour market experience and its square,[12] a time trend and a continuous measure of the local unemployment rate in the given year. Our set of time-invariant characteristics include measured cognitive ability (as proxied by a standardized test score)[13] highest grade completed by the respondent's mother and father, number of siblings, and whether the respondent reported to be in a broken home as of age 14. The dependent variable used is the reported log hourly wage at the respondent's most recent job, which is converted to real 1993 dollars. Summary statistics are provided in Table II.

As mentioned previously, to gather data where education changes over time for a given individual, we necessarily must look to the wage outcomes of younger workers. As argued above, we first restrict attention to only those individuals who are quite active in the labour force, so that the reported wages are more likely to measure the actual earnings potential of the individuals. Additionally, one might be concerned that our reported estimates of returns to education would be biased simply because individuals could be paid a premium for commitment to full-time employment. That is, if we track the wage outcomes of individuals over time, the change in those wages might not be solely attributable to changes in education (even after controlling for the effect of other variables), simply because upon completion of schooling an individual might be paid some premium for committing to full-time employment or for the potential to work flexible hours. If this story is true, it might suggest that reported wages for young workers that go on to obtain further education are misleading, since they could have earned more than the hourly wage reported had they committed to full-time employment.

[12] Potential experience is defined as Age − Education − 5.

[13] This test score is constructed from the 10 component tests of the Armed Services Vocational Aptitude Battery (ASVAB) which was administered to the NLSY participants in 1980. Since individuals in the sample varied in age at the time of the tests, each of the 10 tests is first residualized on age, and our test score is defined as the first principal component of the standardized residuals. This measure has been employed in previous work, including Cawley *et al*. (1997), Heckman and Vytlacil (2001) and DiNardo and Tobias (2001).

Table II. Summary statistics

Variable	Mean	Std. dev.	Minimum	Maximum
Experience	8.36	4.13	0	22
Unemp. rate	7.68	4.36	0.100	81.0
Education	12.68	1.92	9.00	20
Log wage	2.30	0.053	0.067	4.57
Ability	0.239	0.840	−2.90	2.01
Momed	12.56	1.88	9	20
Daded	13.17	2.56	9	20
Broken	0.157	0.364	0	1
Numsibs	2.83	1.81	0	15

Note: The time-invariant statistics are based on a smaller sample of 1694 individuals for a total of 14,170 person-year observations. In this smaller sample, we restrict attention to those individuals whose parents have at least 9 years of schooling.

To investigate this issue rather broadly, we consider the wage outcomes of white males aged 16–18 in 1981. We also examine the educational outcomes of these same individuals in 1993, and thus are able to determine if these individuals ever go on to complete high school or take some form of college education. We then nonparametrically compare the 1981 wage distribution of those who never obtain a high school education to those who ultimately get at least a high school degree. For the sake of comparison, we repeat this exercise and nonparametrically obtain estimates of the hourly wage density for those with and without at least a high school degree in 1993. The basic idea motivating this exercise is that if such a flexibility or commitment premium exists, then dropouts would be more likely to receive this premium in their 1981 wage than those who go on to complete at least high school, since they would be more likely to commit to full-time work.

As Figure 1 shows, the 1981 nonparametric density estimates are rather similar for both groups, suggesting (as we might expect) that in 1981, we have a sample of primarily fixed hourly wage earners, and those that never obtain at least a high school degree are not clearly receiving a higher hourly wage than those who ultimately do decide to complete at least this degree. This provides some suggestive evidence that the hourly wages reported by these young workers might reasonably reflect their true earnings potential at the given age and schooling level. The 1993 density estimates tell a different, and certainly expected story. The wage distribution for workers with at least a high school degree is clearly shifted to the right relative to those without this degree.

It is also worthwhile to discuss the nature of time-variation in educational attainment in our final sample. Sixty-two percent of the individuals whose education values change over the sample period experience only one change in the quantity of schooling attained during this period. Thus, a 'typical' observation might be an individual with 11 years of schooling who then completes his high school education, and is observed throughout the rest of the sample with 12 years of education. Some individuals in our sample do report more than one change in the quantity of schooling attained, but 97.5% of them have three or fewer changes over the sample period. While our sample selection mechanism does not preclude the possibility of tracking wage changes for individuals prior to completing high school (e.g. following wage growth from, say, 10 to 11 years of education), these types of changes are relatively infrequent. In fact, 92.2% of the changes in educational attainment in our final sample are changes resulting from the completion of a high school degree or changes representing the completion of some years of post-secondary education.

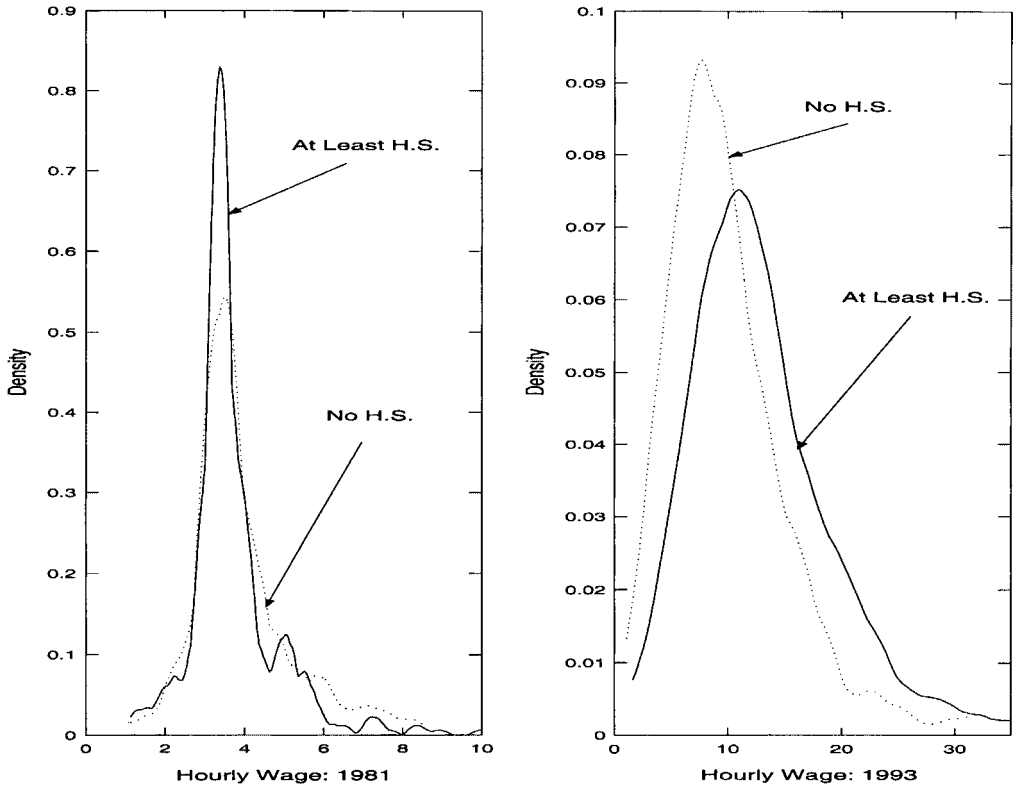

Figure 1. Nonparametric estimates of hourly wage density for white males who obtain at least a high school (H.S.) degree and those who do not: 1981 and 1993

To get some idea about the nature of possible vocational changes resulting from changes in completed years of schooling, we went back to the NLSY and extracted the census three digit industry and occupation codes associated with the respondent's most recent job. When a change in education occurred, we checked to determine if the increase in years of schooling completed was associated with a change in the occupation or industry code at the respondent's most recent job. We found that 76% of all changes in highest grade completed were met by changes in the census occupation code, and 64% of the educational changes were associated with changes in industry. This finding lends additional support to the notion that wage changes reported upon the completion of additional schooling may be a reasonable reflection of earnings potential at that level of schooling, given that individuals are often switching occupations and industries upon acquiring more education.

5. EMPIRICAL RESULTS

The set of competing models outlined in the previous section are estimated using the Gibbs sampler described in Appendix A. In our Gibbs runs we take 11,000 replications. We discard the first 1000 to mitigate start up effects and use the remaining 10,000 to calculate our posterior features of

interest. Diagnostics such as numerical standard errors indicate a high degree of accuracy with this number of runs, and the use of *blocking* or *grouping* steps (e.g. Chib and Carlin, 1999) also helps to mitigate autocorrelation in the chains. The priors are described in Appendix B. Suffice it to note here that we centre our prior over parameter values that seem to us reasonable, and we then choose large values for prior covariance matrices (or small values for degrees of freedom parameters) so as to ensure that the priors are quite noninformative relative to the data. We also check the effect of our prior by comparing results with this relatively noninformative prior to a fully noninformative prior (which is still proper for Σ to ensure propriety of the posterior). Results for our two priors are virtually the same, indicating that data information is predominant.

5.1. Are Returns to Schooling Heterogeneous?

An important focus of this paper is whether heterogeneity exists in the returns to schooling relationship and, if so, what is the best way to model that heterogeneity. As to the existence question, note that our formal comparison of Model 3 against Model 2 tests if the data prefer the added flexibility of permitting heterogeneity in returns to education. To formally compare these models, we calculate *log marginal likelihoods*. Specifically, we note

$$\underbrace{\frac{\Pr(M_i|y)}{\Pr(M_j|y)}}_{\text{Posterior Odds}} = \underbrace{\left(\frac{p(y|M_i)}{p(y|M_j)}\right)}_{\text{Bayes Factor}} * \underbrace{\left(\frac{p(M_i)}{p(M_j)}\right)}_{\text{Prior Odds}}$$

where $p(y|M_i)/p(y|M_j)$, the *Bayes factor*, is the ratio of the *marginal likelihoods* from model M_i to model M_j. Under equal prior odds (i.e. $p(M_i) = p(M_j)$), the posterior odds comparing Model i to Model j simply reduces to the ratio of marginal likelihoods.

Using the method described by Chib (1995) we find that the log marginal likelihoods associated with Models 1–3 were 12,413, −8046, and −8153, respectively. Posterior odds ratios can be obtained by differencing these log marginal likelihoods, and then exponentiating the resulting difference. Thus, the data *overwhelmingly* prefer the bivariate normal heterogeneity Model 2 over the heterogeneity in intercepts only Model 3 and the no-heterogeneity Model 1.

As for the second part of this question, we seek to determine if bivariate normality provides an adequate description of the heterogeneity. To this end we calculate the Bayesian Information Criterion (BIC) for each of our alternate models using our noninformative prior. Use of the BIC not only enables us to see if our results are robust to different model selection criteria, but also enables us to determine if our ordering of Models 1–3 is sensitive to the fact that informative priors were required in those calculations. We compute BIC as follows:

$$\text{BIC}_j = 2 \log p(y|\theta = \hat{\theta}_j) - p_j \log n$$

where $\hat{\theta}_j$ denotes the maximum likelihood estimate for model j and p_j denotes the number of parameters in model j.[14]

Regardless of whether one prefers marginal likelihoods (calculated using a subjectively elicited informative prior) or information criteria such as BIC (calculated using a noninformative prior), the

[14] Schwarz (1978) shows that $\exp[(\text{BIC}_i - \text{BIC}_j)/2]$ can be used as an approximation to the posterior odds ratio.

results in Table III are overwhelming. Returns to schooling are heterogeneous, and the continuous, normally distributed heterogeneity of Model 2 is massively preferred by the data. The second preferred choice is the two-component normal mixture (Model 5), followed by Model 3 which permits normal heterogeneity in intercepts only. The discrete heterogeneity model ranks fourth among our five competing models, and seems to provide a clearly inferior description of the heterogeneity. Also note that this discrete heterogeneity model is, perhaps, the case where the use of instrumental variables is most promising, and such a discrete heterogeneity assumption has been made in previous work on this topic (e.g. Ichino and Winter-Ebmer, 1999).

In Table IV, we present estimation results of key first and second-stage coefficients from Models 1, 2 and 4. For Model 4, we present results for the $G = 2$ case, which is not the number of components most preferred by the data. However, presenting results for $G = 10$ components would consume a lot of space, and since the basic conclusions can be illustrated by providing the $G = 2$ results, we present only the latter.

Point estimates of most parameters, particularly the γ parameters associated with the time-invariant characteristics, tend not to change across the alternate specifications. We find strong evidence of a quadratic (concave) experience profile, and that the local unemployment rate has a consistently negative impact on log wages. Somewhat surprisingly, we also find a consistently negative estimate associated with our time trend, suggesting that over this period of study real log wages tended to decline by about 2% each year. We interpret this result, however, with caution as the time trend is highly correlated with our potential experience measure (with a correlation

Table III. Model comparison results

	Model 1	Model 2	Model 3	Model 4	Model 5
	$\alpha_i = \alpha$ $\beta_i = \beta$	$\theta_i \sim N(\theta_0, \Sigma)$	$\alpha_i \sim N(\alpha, \sigma_\alpha^2)$ $\beta_i = \beta$	$\theta_i = \theta_g^0$ with prob π_g $G = 10$	Normal mix $G = 2$
BIC	−24,412	−15,866	−16,051	−16,528	−15,898

Note: For Model 4, we report BIC results using $G = 10$, which we found to be the number of support points favoured by the data. For Model 5, we consider the two-component normal mixture model.

Table IV. Posterior means (and std. dev.'s) of key first and second-stage parameters: Models 1, 2 and 4

	Model 1	Model 2	Model 4 ($G = 2$)	
First-stage parameters				
Experience	0.105	0.126	0.112	
	(0.002)	(0.005)	(0.004)	
Experience^2	−0.003	−0.004	−0.004	
	(0.0002)	(0.0001)	(0.0002)	
Time	−0.022	−0.024	−0.017	
	(0.002)	(0.004)	(0.003)	
Unemp. rate	−0.010	−0.004	−0.005	
	(0.001)	(0.001)	(0.001)	
Second-stage parameters			1st Comp. (θ_1^0)	2nd Comp. (θ_2^0)
Intercept	0.574	0.312	0.390	0.605
	(0.034)	(0.075)	(0.044)	(0.046)
Education	0.105	0.114	0.088	0.116
	(0.002)	(0.006)	(0.003)	(0.004)

coefficient equal to 0.77). Expected log wages tend to increase for an individual in their next year of the panel, as the potential experience effect washes out the negative effect of the time trend.[15]

5.2. The Extent of Heterogeneity and the Appropriateness of Normality

Of course, one can get a feeling for the extent of (and thus need to allow for) heterogeneity in returns to schooling without appealing to the marginal likelihood calculations in the previous section. This can be accomplished by simply examining the parameter estimates obtained from Model 2, and in particular, the parameter estimates of the second-stage covariance matrix Σ. We note that, in Model 2, the posterior means of the variance parameters in Σ are: $E(\Sigma_{11}|Data) = 0.9716$ and $E(\Sigma_{22}|Data) = 0.006$, and the marginal posterior distributions of these elements are concentrated in regions away from 0 (i.e. *the posterior allocates virtually no probability to regions where α_i and β_i are constant over individuals*). To see this more clearly, we note that the posterior standard deviation of Σ_{11} was 0.140 and the posterior standard deviation of Σ_{22} was 0.001, which are quite small relative to their mean values.

To gain a feeling for the economic interpretation of these second-stage parameters, note that the point estimates of parameters for Model 2 (see Table IV plus remember that $E(\Sigma_{22}|Data) = 0.006$) imply the individual-level return to schooling parameter $\beta_i \sim N$ (0.114, 0.006). This implies a 95% probability interval would be roughly [−0.04, 0.27], indicating that an added year of schooling increases hourly wages from −4% per year to 27% per year. This is a large amount of heterogeneity, which again provides strong evidence that returns to schooling vary across individuals, as suggested by our formal tests in the previous section. Similar calculations show a large amount of heterogeneity across individuals through baseline differences in intercepts, as a 95% posterior probability interval for the second-stage intercept distribution is [−1.62, 2.24].

To provide some further insight on the appropriateness of bivariate normality, we plot kernel-smoothed densities of the intercept and slope posterior means for every individual in the sample (i.e. $E(\alpha_i|Data)$ and $E(\beta_i|Data)$ for $i = 1, \ldots, N$) using Model 2. These are presented in Figures 2 and 3. Though the figures do not look exactly normal—particularly for the slope distribution—it does seem that normality can provide a reasonable approximation. Though Models 4 and 5 are more flexible than Model 2, and do perform better at approximating the shape of the heterogeneity distribution, our BIC calculations show that any improvements in fit associated with adopting these models are outweighed by penalties associated with added parameterization. Finally, note that we have both 'tested up' and 'tested down' in that we have compared our preferred normal specification to both more restrictive and more general models.

5.3. Limitations and Comparison with the Literature

It is also very interesting to note that our 95% probability interval for returns to schooling 'covers' the majority of point estimates reported in previous studies. For example, the point estimates from all but one[16] of the IV studies reviewed by Card (2001, table II, pp. 1146–1148) fall within our 95% probability interval. This suggests that our direct approach for quantifying the nature of

[15] We also treated the time effect in a nonparametric fashion by adding dummy variables for the various years of the sample. Substantive results were unaffected by this alternate specification.

[16] The exception is the study by Ichino and Winter-Ebmer (1999) who use a dummy for completion of high school as their schooling measure.

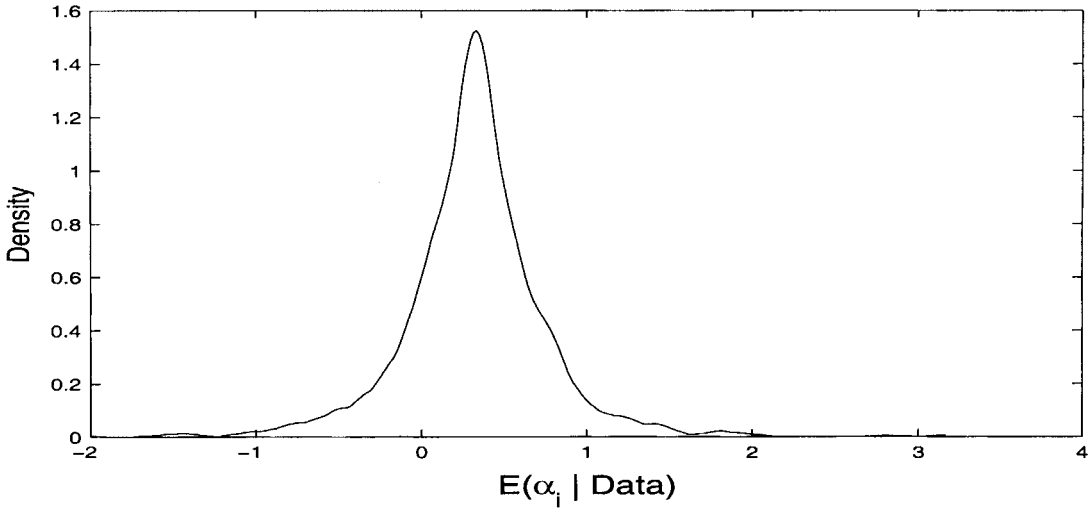

Figure 2. Density of posterior means of α_i parameters

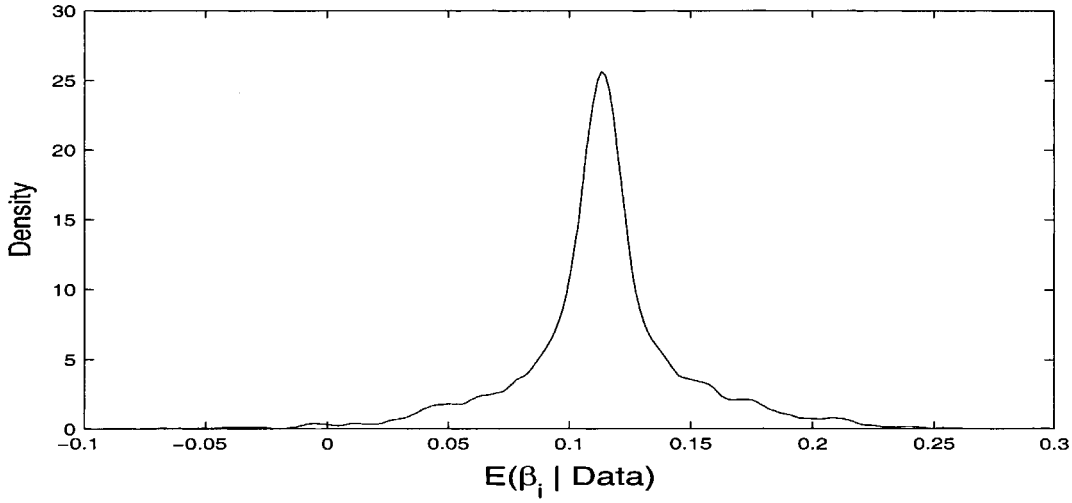

Figure 3. Density of posterior means of β_i parameters

heterogeneity in returns to schooling may be capable of reconciling the disparity of IV estimates produced in previous work.

Though our approach shows promise for reconciling previous results, it is important to interpret this claim with some degree of caution. First, we reiterate that our model maintains the widely used assumption of linearity in education, and some studies have questioned the appropriateness of this assumption. For example, the Becker–Card schooling model (e.g. Card, 1999) contains a quadratic term in education, and numerous other studies have examined the empirical importance of

nonlinearities in education upon degree completion.[17] As such, it is possible that misspecification could affect our estimation of the second-stage heterogeneity distribution. In this paper, we use the popular Mincerian log wage equation with a linear schooling term as our point of departure, and defer issues of nonlinearity in education as the subject of future work.

Second, it is worth noting that our analysis exploits time-variation in schooling for relatively young workers at reasonably low levels of schooling. To the extent that the nature of the heterogeneity distribution differs by age and labour market experience, the comparison between our results and those obtained in IV studies may not be appropriate. This is particularly true since studies employing instrumental variables strategies tend to focus on a sample of older workers who have completed their schooling. Our results may not generalize completely to analyses of these types of workers, given the nature of the NLSY sample. It is worth noting, however, that in the final years of our sample, we do obtain some observations on workers with a reasonable amount of experience. For example, in 1993 workers range in age from 28–36, and thus are very likely to have completed their education and possess a reasonable degree of labour market experience.

Finally, it is useful to review some benefits associated with IV studies that may not be shared by our analysis. Use of IV enables the researcher to estimate a structural 'causal' effect of education on earnings. In the case of a binary schooling–binary instruments model, the IV estimate can be interpreted as the return to education for those individuals who change their schooling status as a 'result' of the instrument. This defines a structural Local Average Treatment Effect (LATE) parameter (e.g. Imbens and Angrist, 1999) that may or may not be the object of primary interest. If it is the object of interest, then IV provides a clean and far more direct route for recovering the parameter than is provided by the methods described in this paper. On the other hand, it is not obvious that the candidate instruments we use in practice are valid, or even if they are valid, that their use generates structural parameters we find interesting. Our analysis attempts to directly characterize the distribution of heterogeneity in the population, which can then be used to calculate features of interest. It is also worth noting that we have assumed the endogeneity of schooling problem can be handled adequately through the incorporation of the individual effects, while IV methods handle the endogeneity issue directly.[18]

5.4. Can Time-invariant Observables Help to Explain the Unobserved Heterogeneity?

All of our previous models did not add any observable, time-invariant characteristics to the second stage of our model in an attempt to explain differences in returns to schooling across individuals. To this end we estimate Model 2W. This model takes our preferred normal heterogeneity Model 2 and adds to it a set of time-invariant ability and family background characteristics.

Table V contains posterior means and standard deviations associated with these second-stage parameters. Also provided in Table V are Bayes factors in favour of each coefficient being equal to zero. A common rule of thumb (see, e.g., Poirier, 1995, p. 380) uses Bayes factors of less than 0.10 to indicate strong evidence (and values between 0.10 and 1.0 to indicate slight evidence) against the hypothesis that the relevant coefficient equals zero.

[17] Two studies that look at nonlinearities and heterogeneity simultaneously are Harmon and Walker (1999) and Arias *et al*. (2001).

[18] If a good instrument is available, one could modify the structural relationship in (3) to account for the endogeneity problem and to simultaneously pursue a 'direct' modelling of the heterogeneity distribution.

Table V. Posterior means of coefficients on W (st. dev.'s in parentheses)

	Model 2W			
	Heterogeneity in intercept	Bayes factor for no effect	Heterogeneity slope	Bayes factor for no effect
Intercept	0.797	—	0.070	—
	(0.408)		(0.031)	
Ability	−0.073	0.270	0.0125	0.001
	(0.079)		(0.006)	
Momed	0.021	3.90	−0.001	86.4
	(0.037)		(0.003)	
Daded	−0.022	2.01	0.002	43.0
	(0.029)		(0.002)	
Broken	0.115	0.610	−0.015	1.07
	(0.155)		(0.013)	
Numsibs	−0.079	0.000	0.007	0.000
	(0.034)		(0.003)	

Note: These results are obtained using a restricted sample of 14,170 observations where parental education was restricted to be at least 9 years.

Generally, the results from Model 2W suggest modest roles for these explanatory variables in the second stage of the hierarchy. We find little evidence that parental education plays any significant role in explaining variation in either intercepts or slopes. However, we do find evidence that both ability and family size (number of siblings) play an important role in these second stage equations. Interestingly, the effect of ability on wages is positive throughout the education support and increases as the individual acquires more education. Similarly, we find a negative effect of family size at small values of education ($Ed < 12$), virtually zero effect at 12 years, and a positive effect at higher values of education.

Despite the significance of ability and number of siblings in the second stage, the addition of the explanatory variables does little to reduce the amount of unobserved heterogeneity. To see this last point more clearly, note that the variance parameters in the covariance matrix Σ can be interpreted as reflecting the extent of heterogeneity that cannot be explained in terms of observables. In Model 2 with this set of data we find $E(\Sigma_{11}|Data) = 1.242$ and $E(\Sigma_{22}|Data) = 0.0080$. With Model 2W we have $E(\Sigma_{11}|Data) = 1.194$ and $E(\Sigma_{22}|Data) = 0.0076$. In other words, the addition of the time-invariant characteristics reduced the second stage intercept variability by only 3.86%, and reduced the unobserved variability in returns to schooling by only 5%. *Thus, we see very little reduction in the total amount of variation across individuals after including these time-invariant characteristics, and conclude that the role of unobserved heterogeneity remains substantial.*

6. CONCLUSIONS AND EXTENSIONS

In this paper, we revisited the issue of identifying and characterizing the extent of unobserved heterogeneity in returns to schooling. Motivated by the recent discussion in Card (2001), we introduced a class of models permitting individual-specific intercepts and slopes. This class can be theoretically justified based on an equilibrium model accounting for forces of both supply and demand. Our econometric approach explicitly models heterogeneity and thus differs from conventional methods which utilize instrumental variables or natural experiments. Our view is

that the approach employed in this paper gives a better way to test for the existence of such heterogeneity, as well as to determine how to best model the distribution of that heterogeneity.

Motivated by the assumptions made or specifications employed in previous studies, we brought different forms of heterogeneity to the data and determined which form was most supported. Our results strongly suggested that returns to schooling were heterogeneous, that discrete distributions for the heterogeneity provide an inferior description of the heterogeneity, and that the simple bivariate normal model does provide a very adequate description.

APPENDIX A: COMPUTATIONAL METHODS

Methods for carrying out Bayesian computation for our alternate models are given in this section. As described in Section 2, all models are special cases of the following general structure:

$$y_{it}|x_{it}, z_{it}, w_i, \theta_i, \gamma, \sigma_\varepsilon^2 \overset{ind}{\sim} N(x_{it}\theta_i + z_{it}\gamma, \sigma_\varepsilon^2), \quad i = 1, 2, \cdots, n, \ t = 1, 2, \ldots, T_i \tag{A.1}$$

$$\theta_i|\lambda, w_i \overset{ind}{\sim} f(\theta_i|\lambda, w_i) \tag{A.2}$$

$$\gamma|\underline{\mu}_\gamma, \underline{V}_\gamma \sim N(\underline{\mu}_\gamma, \underline{V}_\gamma) \tag{A.3}$$

$$\lambda|\underline{\lambda} \sim g(\underline{\lambda}) \tag{A.4}$$

$$\sigma_\varepsilon^{-2}|s_\varepsilon^{-2}, \underline{\nu}_\varepsilon \sim G(\underline{s}_\varepsilon^{-2}, \underline{\nu}_\varepsilon) \tag{A.5}$$

where $f(\theta_i|\lambda, w_i)$ is a hierarchical prior which depends on parameter vector λ and a $1 \times k_w$ vector of explanatory variables w_i, $G(\underline{s}_\varepsilon^{-2}, \underline{\nu}_\varepsilon)$ denotes the Gamma distribution with mean $\underline{s}_\varepsilon^{-2}$ and degrees of freedom $\underline{\nu}_\varepsilon$ (see Poirier, 1995, p. 100)

$$\theta_i = \begin{bmatrix} \alpha_i \\ \beta_i \end{bmatrix}, \quad x_{it} = [1 \quad s_{it}]$$

and z_{it} is a $1 \times k_z$ row vector. Define $k \equiv k_z + 2$, $NT \equiv \sum_{i=1}^N T_i$ and stack all observations using the standard notation. For instance, $\theta = (\theta_1, \ldots, \theta_N)'$

$$X_i = \begin{bmatrix} x_{i1} \\ \cdot \\ \cdot \\ x_{iT_i} \end{bmatrix}$$

and define y_i and Z_i conformably. Define

$$X = \begin{bmatrix} X_1 \\ \cdot \\ \cdot \\ X_N \end{bmatrix}$$

and y, W and Z conformably. Denote all the data, y, X, Z and W as *Data*. As general notation, let Γ denote all the parameters (including, where relevant, latent data) in the model, and Γ_{-x} denote all parameters other than x.

We present in this appendix our posterior simulator for fitting Models 4 and 5. We focus on these two models as the algorithms used for fitting the remaining models are basically special

cases of those used to fit Models 4 and 5. Marginal likelihoods are calculated using the method described by Chib (1995). Complete details regarding the posterior simulator for all the candidate models as well as a deeper discussion of the marginal likelihood calculations can be obtained from: http://orion.uci.edu/~jtobias/ktappend.pdf.

Models 4: $\theta_i = \theta_g^0$ with Probability π_g for $g = 1, \ldots, G$

This varies from a standard mixture model (e.g. Chib, 1995, pp. 1318–1320) due to the fact that we are assuming γ and σ_ε^2 be the same for all individuals. Define the component label vector as $c_i \equiv (c_{1i}, \ldots, c_{Gi})'$ with $c_{gi} = 1$ denoting that the ith observation is drawn from the gth component and let C be the $NG \times 1$ vector stacking c_i for all individuals. Given the component indicators, the conditional likelihood function is

$$p(y|\Gamma) = \prod_{i=1}^{N} [\phi(y_i; X_i\theta_1^0 + Z_i\gamma, \sigma^2)]^{c_{1i}} [\phi(y_i; X_i\theta_2^0 + Z_i\gamma, \sigma^2)]^{c_{2i}} \cdots [\phi(y_i; X_i\theta_G^0 + Z_i\gamma, \sigma^2)]^{c_{Gi}} \tag{A.6}$$

where $\phi(\,)$ denotes the multivariate normal p.d.f. We add the following multinomial hierarchical prior for the component indicators:

$$p(c_i|\pi) = \prod_{g=1}^{G} \pi_g^{c_{gi}} \tag{A.7}$$

where $\pi = (\pi_1, \ldots, \pi_G)'$ with $c_{gi} \in \{0, 1\}$, $0 \leq \pi_g \leq 1$ and $\sum_{g=1}^{G} c_{gi} = \sum_{g=1}^{G} \pi_g = 1$.

The prior for π is given by:

$$\pi \sim Dir(a_1, a_2, , a_G) \tag{A.8}$$

where $Dir(\,)$ denotes the Dirichlet distribution (see Poirier, 1995, p. 132). For γ and σ_ε^{-2} we use priors given by (A.3) and (A.5). Finally, we assume:

$$\theta_g^0 \overset{ind}{\sim} N(\underline{\theta}_g, \underline{V}_g), \quad \forall g = 1, 2, \ldots, G \tag{A.9}$$

Posterior simulation is done using a Gibbs sampler with data augmentation involving the following posterior conditionals. First, for $g = 1, \ldots, G$ we have (e.g. Lindley and Smith, 1972)

$$\theta_g^0|Data, \Gamma_{-\theta_g} \overset{ind}{\sim} N(D_{\theta_g} d_{\theta_g}, D_{\theta_g}) \tag{A.10}$$

where

$$D_{\theta_g} = \left[\sigma_\varepsilon^{-2} \sum_i c_{gi} X_i X_i' + \underline{V}_g^{-1}\right]^{-1}, \quad d_{\theta_g} = \sigma_\varepsilon^{-2} \sum_i c_{gi} X_i'(y_i - Z_i\gamma) + \underline{V}_g^{-1}\underline{\theta}_g$$

Next

$$\sigma_\varepsilon^{-2}|Data, \Gamma_{-\sigma_\varepsilon^{-2}} \sim G(\overline{s}_\varepsilon^{-2}, \overline{\nu}_\varepsilon) \tag{A.11}$$

where

$$\overline{\nu}_\varepsilon = NT + \underline{\nu}_\varepsilon$$

and

$$\overline{s}_\varepsilon^2 = \frac{\sum_{g=1}^{G}\sum_{i=1}^{N}\sum_{t=1}^{T} c_{gi}(y_{it} - x_{it}\theta_g^0 - z_{it}\gamma)^2 + \underline{\nu}_\varepsilon \underline{s}_\varepsilon^2}{\overline{\nu}_\varepsilon}$$

We then have

$$\gamma | Data, \Gamma_{-\gamma} \sim N(D_\gamma d_\gamma, D_\gamma) \tag{A.12}$$

where

$$D_\gamma = \left(\sigma_\varepsilon^{-2}\sum_i Z_i' Z_i + \underline{V}_\gamma^{-1}\right)^{-1}, \quad d_\gamma = \sigma_\varepsilon^{-2}\sum_{g=1}^{G}\sum_{i=1}^{N} c_{gi} Z_i'(y_i - X_i\theta_g^0) + \underline{V}_\gamma^{-1}\underline{\mu}_\gamma$$

Next we have, for $i = 1, \ldots, N$

$$c_i | Data, \Gamma_{-c_i} \overset{ind}{\sim} Mult\left(1, \left[\frac{\pi_1 \phi(y_i; X_i\theta_1^0 + Z_i\gamma, \sigma_\varepsilon^2 I_{T_i})}{\sum_{g=1}^{G}\pi_g \phi(y_i; X_i\theta_g^0 + Z_i\gamma, \sigma_\varepsilon^2 I_{T_i})}, \ldots, \frac{\pi_G \phi(y_i; X_i\theta_G^0 + Z_i\gamma, \sigma_\varepsilon^2 I_{T_i})}{\sum_{g=1}^{G}\pi_g \phi(y_i; X_i\theta_g^0 + Z_i\gamma, \sigma_\varepsilon^2 I_{T_i})}\right]'\right) \tag{A.13}$$

where $Mult(\,)$ denotes the multinomial distribution (see Poirier, 1995, pp. 118–119). Finally, we have

$$\pi | Data, \Gamma_{-\pi} \sim Dir(n_1 + a_1, n_2 + a_2, \ldots n_G + a_G) \tag{A.14}$$

where $n_g = \sum_{i=1}^{N} c_{gi}$.

Posterior simulation proceeds by sequentially simulating from (A.10), (A.11), (A.12), (A.13) and (A.14).

Model 5: $f(\theta_i | \{\pi_g\}, \{\beta_g^0\}, \{\Sigma_g\}) = \sum_{g=1}^{G} \pi_g \phi(\theta_i; \theta_g^0, \Sigma_g)$

Inference for this model follows similarly to that of Model 4 with slight modifications. For Model 5, the conditional likelihood function is:

$$p(y|\Gamma) = \prod_{i=1}^{n} \phi(y_i; X_i\theta_i + Z_i\gamma, \sigma^2) \tag{A.15}$$

Augmenting with component label vectors $\{c_i\}$ as with Model 4, the second stage becomes:

$$\theta_i | \Gamma_{-\theta_i} = [\phi(\theta_i; \theta_1^0, \Sigma_1)]^{c_{1i}} \ldots [\phi(\theta_i; \theta_G^0, \Sigma_G)]^{c_{Gi}} \tag{A.16}$$

and we choose priors for $\{\theta_g^0\}$, $c_i|\pi$ and π as in (A.9), (A.7) and (A.8), respectively. Finally, we put conjugate priors on the inverse covariance matrices Σ_g^{-1}

$$\Sigma_g^{-1} \overset{iid}{\sim} W([\underline{\rho}_g \underline{\Sigma}_g]^{-1}, \underline{\rho}_g)$$

with W denoting the Wishart density (e.g. Poirier, 1995, p. 136). We obtain the following forms for the complete posterior conditionals:

$$\theta_i|\Gamma_{-\theta_i}, \text{Data} \overset{ind}{\sim} N(D_{\theta_i} d_{\theta_i}, D_{\theta_i}) \tag{A.17}$$

where

$$D_{\theta_i} = \left(X_i' X_i/\sigma^2 + \sum_{g=1}^{G} c_{gi} \Sigma_g^{-1} \right)^{-1} \quad \text{and} \quad d_{\theta_i} = X_i'(y_i - Z_i\gamma)/\sigma^2 + \sum_{g=1}^{G} c_{gi} \Sigma_g^{-1} \theta_g^0$$

and

$$\theta_g^0|\Gamma_{-\theta_g^0}, \text{Data} \overset{ind}{\sim} N(D_{\theta_g} d_{\theta_g}, D_{\theta_g}) \tag{A.18}$$

where

$$D_{\theta_g} = (n_g \Sigma_g^{-1} + \underline{V}_g^{-1})^{-1} \quad \text{and} \quad d_{\theta_g} = \sum_{i=1}^{n} c_{gi} \Sigma_g^{-1} \theta_i + \underline{V}_g^{-1} \underline{\theta}_g$$

Finally

$$\Sigma_g^{-1}|\Gamma_{-\Sigma_g^{-1}}, \text{Data} \overset{ind}{\sim} W \left(\left[\sum_{i=1}^{n} c_{gi} (\theta_i - \theta_g^0)(\theta_i - \theta_g^0)' + \underline{\rho}_g \underline{\Sigma}_g \right]^{-1}, n_g + \underline{\rho}_g \right) \tag{A.19}$$

The posterior conditionals for the remaining parameters $\{c_i\}, \pi, \sigma_\in^{-2}$ and γ follow similarly to those described in Model 4.

APPENDIX B: THE PRIORS

We use priors which select reasonable values for prior means, but then we make the priors relatively noninformative by setting prior variances to be large and/or prior degrees of freedom small. As a form of prior sensitivity analysis, we also calculated results for a prior which was fully noninformative (except for the prior on Σ).

For the parameters common to all the models (remaining consistent with the notation employed in Section 2), we set

$$\underline{s}_\varepsilon^2 = 1, \quad \underline{\nu}_\varepsilon = 0, \quad \underline{\mu}_\gamma = 0_{k_z}, \quad \underline{V}_\gamma = I_{k_x}$$

Note that the resulting prior is completely noninformative for the error variance (and the value for $\underline{s}_\varepsilon^2$ is irrelevant)[19] and quite noninformative for γ. The fully noninformative variant of this prior sets $\underline{V}_\gamma^{-1} = 0_{k_z}$.

The prior for Model 1 is completed by setting

$$E \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} 1.0 \\ 0.1 \end{pmatrix}, \quad \text{Var} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} 1.0 & 0 \\ 0 & 0.1 \end{pmatrix} \tag{B.1}$$

[19] The use of improper prior over nuisance parameters which appear in all models is a common practice. We follow the standard practice of assuming the integrating constant for such noninformative priors over nuisance parameters is the same in all models and, thus, will cancel out in Bayes factor calculations and can be ignored.

These hyperparameter values (which we also use for comparable parameters in other models) are sensible in light of previous empirical work. All coefficients are assumed, *a priori*, to be independent of one another. This defines the hyperparameters for Model 1.

The prior for Model 2 is completed by setting

$$\underline{\rho} = 3 \quad \text{and} \quad \underline{\Sigma} = \begin{pmatrix} 1.0 & 0 \\ 0 & 0.1 \end{pmatrix}$$

and choosing values for $\underline{\theta}$ and $\underline{V}_\theta$. When the second stage of the hierarchy does not depend on explanatory variables w we set:

$$\underline{\theta} = \begin{pmatrix} 1.0 \\ 0.1 \end{pmatrix} \quad \text{and} \quad \underline{V}_\theta = \begin{pmatrix} 1.0 & 0 \\ 0 & 0.1 \end{pmatrix}$$

Motivation for these choices is similar to those for Model 1. That is, we are centring mean effects in regions suggested by our study of the literature and allowing for a moderate degree of heterogeneity, but prior variances and degrees of freedom are selected so as to imply a prior which is noninformative relative to the data. The fully noninformative variant of this prior sets $\underline{V}_\theta^{-1} = 0$. Note, however, that a proper prior for Σ^{-1} is required since an improper prior can lead to an improper posterior (see, e.g., Hobert and Casella, 1996). Hence, our 'fully noninformative' prior is still (weakly) informative about Σ. We note in passing that we have carried out a prior sensitivity analysis with respect to the prior for Σ. Multiplying $\underline{\Sigma}$ by 0.01 or 100 does not substantively alter our results. For the sake of brevity, we do not present results for this prior sensitivity analysis.

When the second stage of the hierarchy does depend on explanatory variables, we choose the prior mean and variance for the 1st and $(k_w + 1)$st elements of θ_0 as given in our prior parameters $\underline{\theta}$ and $\underline{V}_\theta$ above. All other prior means were set to zero, all other prior variances set to 1.0 and all prior covariances set to zero. In other words, when these explanatory variables are included, we continue to centre the prior over Model 2.

To complete the prior for Model 4, we specify $\underline{\theta}_g$ and $\underline{V}_g$ as in (B.1), for all g. For the component probabilities, we make the noninformative choice of $a_1 = a_2 = \cdots = a_G = 1$.

For Model 5, we use a prior that combines the ideas of Models 2 and 4. Thus, priors for θ_g^0 and Σ_g^{-1} for $g = 1, \ldots, G$ are independent of one another, each identical to those used for θ and Σ^{-1} in Model 2. The prior for π in Models 4 and 5 is identical. Finally, for both Models 4 and 5, we impose the identifying restriction $\beta_1^0 < \beta_2^0 < \cdots < \beta_G^0$ through the prior.

ACKNOWLEDGEMENTS

We would like to thank Jun Ishii, Mingliang Li, Peter Schmidt, Jeff Wooldridge and seminar participants at the 2002 Econometric Society European Meetings, the 2002 American Statistical Association Meetings, Brunel University, Claremont-McKenna College, Indiana University, LSU, Michigan State University, Queen Mary and Westfield University, Rice, Texas A&M, UC-Riverside and York University for helpful comments and suggestions. All errors are, of course, our own.

REFERENCES

Allenby GM, Rossi P. 1999. Marketing models of consumer heterogeneity. *Journal of Econometrics* **89**: 57–78.

Allenby G, Arora N, Ginter J. 1998. On the heterogeneity of demand. *Journal of Marketing Research*, forthcoming.
Angrist JD, Krueger AB. 1991. Does compulsory school attendance affect schooling and earnings? *Quaterly Journal of Economics* **106**: 979–1014.
Angrist J, Newey WK. 1991. Over-identification tests in earnings functions with fixed effects. *Journal of Business and Economic Statistics* **9**(3): 317–323.
Arias O, Hallock K, Sosa-Escudero W. 2001. Individual heterogeneity in the returns to schooling: instrumental variables quantile regression using twins data. *Empirical Economics* **26**: 7–40.
Ashenfelter O, Mooney JD. 1968. Graduate education, ability and earnings. *Review of Economics and Statistics* **50**(1): 78–86.
Becker G, Chiswick B. 1966. Education and the distribution of earnings. *American Economic Review* **56**: 358–369.
Belman D, Heywood J. 1991. Sheepskin effects in the returns to education: an examination of women and minorities. *Review of Economics and Statistics* **73**: 720–724.
Blackburn M, Neumark D. 1992. Unobserved ability, efficiency wages and interindustry wage differentials. *Quarterly Journal of Economics*: 1421–1436.
Blackburn M, Neumark D. 1993. Omitted-ability bias and the increase in the return to schooling. *Journal of Labor Economics* **11**: 521–544.
Blackburn M, Neumark D. 1995. Are OLS estimates of the return to schooling biased downward? Another look. *Review of Economics and Statistics* **77**: 217–230.
Carlin B, Polson N. 1991. Inference for nonconjugate Bayesian models using the Gibbs sampler. *Canadian Journal of Statistics* **19**: 399–405.
Card D. 1999. The causal effect of education on earnings. In *Handbook of Labor Economics*, Vol. 3A, Ashenfelter OC, Card D (eds). North-Holland: Amsterdam; 1801–1863.
Card D. 2001. Estimating the return to schooling: progress on some persistent econometric problems. *Econometrica* **69**: 1127–1160.
Cawley J, Conneely K, Heckman J, Vytlacil E. 1997. Cognitive ability, wages and meritocracy. In *Intelligence, Genes and Success: Scientists Respond to the Bell Curve*, Devlin B, Feinberg SE, Resnick DP, Roeder K (eds). Springer: New York; 179–192.
Chib S. 1995. Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association* **90**: 1313–1321.
Chib S, Carlin B. 1999. On MCMC sampling in hierarchical longitudinal data models. *Statistics and Computing* **65**: 361–394.
Chiswick B. 1974. *Income Inequality: Regional Analyses Within A Human Capital Framework*. Columbia University Press: New York.
DiNardo J, Tobias J. 2001. Nonparametric density and regression estimation. *Journal of Economic Perspectives* **15**: 11–28.
Geweke J. 1993. Bayesian treatment of the independent Student t linear model. *Journal of Applied Econometrics* **8**: 19–40.
Griliches Z. 1979. Sibling models and data in economics: beginnings of a survey. *Journal of Political Economy*: S37–S64.
Grogger J, Eide E. 1995. Changes in college skills and the rise in the college wage premium. *Journal of Human Resources* **30**: 280–310.
Hansen WL, Weisbrod BA, Scanlon W. 1970. Schooling and earnings of low achievers. *American Economic Review* **60**: 409–418.
Harmon C, Walker I. 1999. The marginal and average returns to schooling in the UK. *European Economic Review* **43**: 879–887.
Hause J. 1972. Earnings profile: ability and schooling. *Journal of Political Economy* **80**: S108–S138.
Heckman J, Polachek S. 1974. Empirical evidence on the functional form of the earnings–schooling relationship. *Journal of the American Statistical Association* **69**: 350–354.
Heckman J, Singer B. 1984. A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica* **52**: 271–320.
Heckman J, Vytlacil E. 1998. Instrumental variables methods for the correlated random coefficient model: estimating the rate of return to schooling when the return is correlated with schooling. *Journal of Human Resources* **23**: 974–987.

Heckman J, Vytlacil E. 2001. Identifying the role of cognitive ability in explaining the level and change in the return to schooling. *Review of Economics and Statistics* **83**: 1–12.
Heckman J, Layne-Farrar A, Todd P. 1996. Human capital pricing equations with an application to estimating the effect of schooling quality on earnings. *Review of Economics and Statistics* **78**: 562–610.
Heywood J. 1994. How widespread are sheepskin returns to education in the U.S.? *Economics of Education Review* **13**: 227–234.
Hobert J, Casella G. 1996. The effect of improper priors on Gibbs sampling in hierarchical linear mixed models. *Journal of the American Statistical Association* **91**: 1461–1473.
Hungerford T, Solon G. 1987. Sheepskin effects in returns to education. *Review of Economics and Statistics* **69**: 175–177.
Ichino A, Winter-Ebmer R. 1999. Lower and upper bounds of returns to schooling: an exercise in IV estimation with different instruments. *European Economic Review* **43**: 889–901.
Imbens G, Angrist J. 1994. Identification and estimation of local average treatment effects. *Econometrica* **62**: 467–475.
Jaeger D, Page M. 1996. Degrees matter: new evidence on sheepskin effects in returns to education. *Review of Economics and Statistics* **78**: 733–740.
Kling J. 2001. Interpreting instrumental variables estimates of the returns to schooling. *Journal of Business and Economic Statistics* **19**: 358–364.
Lam D, Schoeni RF. 1993. Effects of family background on earnings and returns to schooling: evidence from Brazil. *Journal of Political Economy* **101**(4): 710–740.
Lindley D, Smith AFM. 1972. Bayes estimates for the linear model. *Journal of the Royal Statistical Society, Series B* **34**: 1–41.
Mincer J. 1974. *Schooling, Experience and Earnings*. Columbia University Press: New York.
Murnane R, Levy F, Willett J. 1995. The growing importance of cognitive skills in wage determination. *Review of Economics and Statistics* **77**: 251–266.
Poirier D. 1995. *Intermediate Statistics and Econometrics*. MIT Press: Cambridge, MA.
Schwarz G. 1978. Estimating the dimension of a model. *Annals of Statistics* **43**: 1481–1490.

