Learning About Unobserved Heterogeneity in Returns to Schooling[1]

March, 2002

Gary Koop
University of Glasgow
Department of Economics
Adam Smith Building
Glasgow G12 8RT
g.koop@socsci.gla.ac.uk

Justin L. Tobias
University of California-Irvine
Department of Economics
3151 Social Science Plaza A
Irvine, CA 92697-5100
jtobias@uci.edu

ABSTRACT: We investigate the existence and nature of individual-level heterogeneity in returns to education using panel data taken from the National Longitudinal Survey of Youth (NLSY). We present operational techniques that do not require natural experiments or instrumental variables, but instead introduce and estimate various hierarchical models which 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. Additionally, we describe and employ methods for predicting the distribution of returns to schooling for various subpopulations, and link our methods to "treatment effect" parameters often used in the program evaluation literature. Empirical results indicate that heterogeneity is present in returns to education. Furthermore, we find strong evidence that the heterogeneity follows a continuous rather than discrete distribution, and that bivariate Normality provides a very reasonable description of individual-level heterogeneity in intercepts and returns to schooling.

[1] Corresponding Author: Justin L. Tobias. We would like to thank Jun Ishii, Mingliang Li, and seminar participants at York University, Brunel University, Queen Mary and Westfield University, Claremont-McKenna College, and UC-Riverside for helpful comments and suggestions. All errors are, of course, our own.

1 Introduction

Recently, considerable attention has been given to the existence of unobserved heterogeneity in returns to schooling. In a review of the literature, Card (2001) surveys recent attempts to identify the causal effect of education on earnings using supply-side instrumental variables. Importantly, he describes conditions under which the widely-used instrumental variable (IV) approach will consistently estimate the average return to education in the population when schooling returns are heterogeneous. He also describes the potential limitations of the instruments often-used in empirical work, such as distance to college and compulsory schooling laws.

While the focus of IV studies is often on consistent estimation of the *average* return to schooling in the population, such methods generally reveal little information about the nature of the underlying heterogeneity. Furthermore, in the presence of heterogeneity in returns to schooling, different instruments essentially define different *treatment effects*, as IV can be interpreted as estimating the return to schooling for those who comply with the assign-to-treatment status generated by the particular instrument (Imbens and Angrist (1994), Angrist, Imbens and Rubin (1996)).[2] While the existence of heterogeneity in IV estimates has been interpreted as evidence of underlying heterogeneity in returns to education (*e.g.* Ichino and Winter-Ebmer (1999)), the use of inappropriate instruments might also give rise to observed point estimates that vary across instrument sets.[3] Furthermore, an examination of cases where IV estimates differ only provides suggestive evidence of heterogeneity, and does not enable one to learn about the precise nature and extent of that heterogeneity.

The existence of heterogeneity in returns to schooling also exacerbates the difficulty of evaluating a given policy (*e.g.* a tuition subsidy), and the output of an IV procedure might not provide any meaningful estimates for those seeking to evaluate a particular intervention (Heckman (1997), Ichino and Winter-Ebmer (1999), Card (2001)). To evaluate future programs or interventions, the average return to schooling in the population - the focus of most IV studies - might not provide estimates of the quantities needed. For IV to be useful in this setting would require a valid instrument that "picks off" or identifies the return for the group that would actually be affected by the given intervention.

[2] 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. Card (2001, pp. 1138-1139) also questions the validity of this particular instrument, since college proximity may affect the mapping between ability and schooling, thus potentially violating conditions required for consistent estimation of the average return to education.

[3] For example, Ichino and Winter-Ebmer (1999) use father's military status and father's education as instruments, which arguably could affect earnings directly, or through channels other than simply the "indirect" effect on the quantity of schooling. Bound, Jaeger and Baker (1995) revisit the pioneering work by Angrist and Krueger (1991) and argue that the large number of weak instruments employed may bias some of the reported results. Bound and Jaeger (1996) also question whether the quarter-of-birth instrument is correlated with unobserved ability, and thus is a valid instrument.

In practice, the availability or successful identification of such an instrument is unlikely, and thus a more structural approach might prove to be beneficial for policy evaluation.

In this paper, we investigate heterogeneity in returns to education in a different way, which does not require the use of instrumental variables or natural experiments. We describe and employ methods for uncovering heterogeneity in returns to schooling given the availability of longitudinal data. 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. In the Bayesian literature, such two-stage models are referred to as *hierarchical* and second stage assumptions are labelled as prior. We adopt such an interpretation here, but it is worth noting that this is just a semantic choice and the non-Bayesian would interpret the second stage as part of the likelihood function. We also estimate generalized 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.

We take panel data from the National Longitudinal Survey of Youth (NLSY) to estimate models which 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 in this setting the number of observations obtained for each individual is rather small, it is not particularly problematic 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 time series information from the outcomes of the given individual and cross-section information obtained from the parameter estimates of all individuals. This potential for "pooling" or "shrinking" estimates across individuals as well as our ability to obtain exact finite sample results suggests that a careful Bayesian hierarchical analysis is particularly well-suited for investigating the existence and extent of heterogeneity in returns to education.

Our view is that the approach described here offers several advantages over previous (typically IV) methods for investigating heterogeneity in returns to education. First, the methods we describe allow us to formally *test* for the presence of heterogeneity in returns to schooling across individuals rather than simply assuming that such heterogeneity is present, or arguing that it exists simply because point estimates differ across choices of instruments. Second, the described methods allow us to bring a variety of competing forms of the heterogeneity to the data, and determine the specifications that are most appropriate. This includes, for example, testing discrete heterogeneity models versus continuous alternatives, as both of these specifications have been either formally or informally assumed in the literature (*e.g.* Ichino and Winter-Ebmer (1999) implicitly assume that the heterogeneity is discretely distributed, while Heckman and Vytlacil (1998) introduce an elaborated model where the heterogeneity is continuously distributed). Finally, we also introduce a methodology that allows one to predict various distributions of returns to schooling in the presence of unobserved heterogeneity, and to assess the impact of hypothetical interventions on various subpopulations. This is related to the recent literature on program evaluation, and we link our predictive distributions to various "treatment effects" such as the Local Average Treatment Effect (LATE) (Imbens and Angrist (1994)), which would measure the return to schooling for a subgroup "complying" with a particular instrument.

The outline of the paper is as follows. In the following section, we present in a general framework our hierarchical model for investigating heterogeneity in returns to schooling. In section 3, we focus on some special cases of this general framework which have appeared in the literature, and discuss the interpretation of each model. We also introduce some elaborations of these basic hierarchical models which might shed some new light on the nature of the heterogeneity across individuals. A description of the data is presented in section 4, 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 prior are included in appendices.

2 Learning About the Nature of Unobserved Heterogeneity

The class of models that we analyze in this paper allow for individual-level variation in the intercept and return to education: [4]

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

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

[4] 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.

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.* labor 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.[5] We also assume, as is often employed in this literature following early work by Mincer (1958), Becker and Chiswick (1966), and Heckman and Polachek (1974) (among many others), that the log-wage is linear in schooling.[6]

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 to best 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 error in schooling have been investigated in recent work. If these problems exist, they would lead to biased and inconsistent estimates of underlying structural parameters of interest.

To address these concerns, 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

[5] 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).)

[6] 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, Layne-Ferrar and Todd (1996), Jager 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.

reduced form one, our view is that direct analysis of this reduced form relationship is of interest for a variety of reasons. First, we note that since many studies in the empirical education literature have directly analyzed 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. Relatedly, we note that some of our analysis also adds controls for potentially important omitted variables such as measured cognitive ability and family characteristics. Hence, variables often argued to be responsible for the endogeneity and omitted variables problem can be controlled for, potentially eliminating the need to analyze a more structural specification.[7] Second, analyzing the joint distribution of schooling quantity and wages typically requires some exclusion restriction or instrument that affects schooling choice without affecting wages, thus taking us back to the problems associated with instrumental variables that we originally intended to avoid. Finally, as we will discuss below, under reasonable assumptions about the form of the structural model, even if the endogeneity problem is not ignorable, the heterogeneity in the reduced form will be of the identical form as that in the structural form. Thus, insofar as interest centers on investigating the existence and form of heterogeneity, it does not matter whether we are working with a reduced or structural form.

To understand this last point, it will prove to be useful to look deeper at the relationship between the reduced form and structural parameters. One such structural model, which permits endogeneity of schooling choice,[8] would be written as follows:

$$\begin{aligned} y_{it} &= \tilde{\alpha} + \tilde{\beta}_i s_{it} + \tilde{\gamma} z_{it} + \epsilon_{it} \qquad (3) \\ s_{it} &= \delta_i + \eta v_{it} + u_{it}. \end{aligned}$$

In the above, $\widetilde{\beta}_i$ is the structural returns to schooling parameter, δ_i is an individual-specific intercept that may capture variation in costs of schooling across individuals and v_{it} a vector of explanatory variables for schooling choice. The individual-specific intercepts affect the quantity of schooling choice (and thus would pick up unobservable heterogeneity in the costs of schooling) and affect wages only indirectly through the quantity of schooling attained. The coefficients in the reduced and structural form will be equivalent if wages and schooling (conditional on the covariates and parameters) are determined independently - that is if ϵ_{it} and u_{it} are independent. However, one might still believe that unobservables making an individual more likely to attain more schooling might also make him or her

[7] 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, Weisbrod and Scanlon (1970), Hause (1971), 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. As mentioned later, Blackburn and Neumark (1995) test for the presence of endogeneity and measurement error in schooling, and find no evidence of it after controlling for test scores in the NLSY data we are using.

[8] Such models which permit endogeneity of schooling have appeared numerous times in the literature - see, for example, Angrist and Krueger (1991, page 997, equations (1) and (2)).

more likely to earn higher wages (i.e. we would expect ϵ_{it} and u_{it} to be positively correlated), making analysis of the joint structural model of interest (assuming interest centers on $\tilde{\alpha}$ and $\tilde{\beta}_i$).

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, z and the parameters). The relationship between structural and reduced forms can be derived by working out the relevant conditional distribution implied by (3). For example, if we make the common assumption that

$$\begin{bmatrix} \epsilon_{it} \\ u_{it} \end{bmatrix} \sim N\left(0, \begin{bmatrix} \sigma_\epsilon^2 & \sigma_{\epsilon u} \\ \sigma_{\epsilon u} & \sigma_u^2 \end{bmatrix}\right),$$

then it follows that

$$\beta_i = \tilde{\beta}_i + \frac{\sigma_{\epsilon u}}{\sigma_u^2}.$$

Hence, the variability in reduced-form slope coefficients across individuals equals the variability of the underlying structural coefficients across individuals.[9] *Thus, analyzing the conditional distribution implied by (3) is sufficient if our primary goal is to learn about the existence and extent of variation in returns to education across individuals.* Since the extent of variability in returns to schooling across individuals is obviously a focus of our study, it is useful to note that we can analyze the reduced form relationship to estimate the extent of that variability.

As for measurement error in schooling, we do not investigate this issue, and simply assume that education is measured correctly. As described in section 4, we are careful to exclude individuals whose education is clearly mis-reported or obviously suspect, which undoubtedly helps to mitigate the problem. Further, some previous work has examined the issue of measurement error in education using the data set we employ here. Using NLSY data, Blackburn and Neumark (1995, page 228) summarize their findings on this issue 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 we also make use of the NLSY and control for measured ability, this suggests that it is not a significant problem to abstract from issues of measurement error or endogeneity in our study.

[9] This also assumes the explanatory variables in the reduced form model includes v_{it} . It is important to recognize that this relationship does not hold for the intercept parameters and the mapping becomes even less clear if we let $\tilde{\alpha} = \tilde{\alpha}_i$. However, our primary focus here is on variation in β_i across individuals.

Thus, we take the model in (1) as our primary model of interest, and seek to learn about the nature of the underlying heterogeneity. To this end, the complete, and most general model we specify is given as follows:

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

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

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

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

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

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 which depends on a parameter vector λ and a $1 \times k_w$ row vector w_i. In most cases, w_i will simply include a constant term. In (4), we have also defined

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

As stated previously, our primary goal is to learn about f and λ, and to that end, we revisit some previous assumptions regarding these quantities that have been employed in the literature.[10]

3 A Set of Competing Models

In this section we discuss the various sorts of heterogeneity we wish to consider. In technical terms, this amounts to a selection of a particular form for f in (5). The choice of f is of most relevance for our empirical results. Other, less important, modeling choices and further technical details are provided in Appendices B and C.

To begin, we consider the model that is arguably the most widely used in this literature - dating to the work of Mincer (1958). This model presumes a common schooling effect across individuals – an assumption which has continued to persist in more recent work (*e.g.* Angrist and Krueger (1991), Ashenfelter and Krueger (1995), Blackburn and Neumark (1995) and Angrist (1995)).[11] Thus, we

[10] For the non-Bayesian reader, note that (6), (7) and (8) specify priors for the second stage parameters. 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, page 100). As λ will change across our models, we represent our prior for these parameters generically by g in (7), and suppose that this prior depends on some hyperparameters $\underline{\lambda}$. In our empirical work, we use relatively noninformative priors and data information is predominant. Thus, the reader wishing to understand the main modeling issues and empirical results of the paper can focus on (4) and (5) as being the key equations.

[11] See, for example, Angrist and Krueger (1991, page 221, equation (2)), Ashenfelter and Krueger (1994, pages 1161 and 1161, equations (1), (2), and (6)), Angrist (1995, page 1072, equation (2)), and Blackburn and Neumark (1995, page 221, equation(2)). Angrist (1995) does permit schooling effects to vary over time as well as across schooling groups.

consider as a baseline case a model that imposes constant intercepts and slopes across individuals:

Model 1: $\alpha_i = \alpha$**,** $\beta_i = \beta$**.**

This most restrictive specification amounts to a simple regression model where everyone is assumed to have the same individual effect (intercept) and marginal return to an added year of schooling. In statistical language, this model specifies the heterogeneity distribution, f, as being degenerate at the point (α, β). By comparing this model to a variety of competitors we can formally *test* if models permitting heterogeneity in returns to schooling are favored by the data. This is quite different than the approach that has been taken by some researchers where it is argued that heterogeneity exists simply because point estimates of the return to schooling differ across different choices of instruments.

We elaborate this baseline model in several ways, and in each model, we permit heterogeneity of a distinct form. In the first extension of this baseline model, we adopt a hierarchical framework to investigate the issue of heterogeneity. This leads us to our second specification:

Model 2: $\theta_i | \theta_0, \Sigma \overset{iid}{\sim} N(\theta_0, \Sigma)$.

In this model, we allow slopes and intercepts to differ across individuals, but allow for a degree of similarity across people by assuming that individuals are drawn from the same Normal population. Marginal posterior distributions of the person-specific intercepts and returns to schooling will then combine information provided by outcomes for that individual with information provided by other individuals that is incorporated in the second stage. Such hierarchical or "random effects" models have been suggested by or employed in past work in the schooling literature by Becker and Chiswick (1966), Chiswick (1974), and Mincer (1974), and more recently theoretical issues in an elaborated version of this model were described in Heckman and Vytlacil (1998).

One can also elaborate Model 2 by proposing that observable and time-invariant variables, such as measured cognitive ability or family endowments, might explain variation in the intercept and return to schooling parameters across individuals (when including these covariates, we denote the model as **Model 2W**, with the W denoting the explanatory variables added to the second stage of the hierarchy). In this way, we can characterize both *observed* and *unobserved* heterogeneity in returns to education across individuals. We incorporate this feature into Model 2 by defining

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

which leads us to:

Model 2W: $\theta_i | w_i, \theta_0, \Sigma \overset{ind}{\sim} N(w_i\theta_0, \Sigma)$.

Note that $w_{\alpha i}$ represents covariates that explain variation in intercepts across individuals, while $w_{\beta i}$ are used to explain variation in returns to schooling across individuals. For example, we might expect that measured cognitive ability shifts the means of the intercept and return to education distributions and thus decide to include it as an element of both $w_{\alpha i}$ and $w_{\beta i}$.[12]

We also note that a comparison of Model 1 versus Model 2 (without the additional covariates) doesn't directly test for the existence of heterogeneity in returns to education. In fact, one might find a preference for Model 2 over Model 1 simply because Model 2 allows for individual-specific intercepts, while the ability to additionally account for heterogeneity in returns to education is not supported by the data. In other words, a preference for Model 2 over Model 1 does not necessarily indicate the existence of heterogeneity in returns to schooling, but rather, simply indicates a need to control for *some form of heterogeneity at the individual-level.*

One way to get at this issue is to simply investigate the extent of heterogeneity in returns to schooling from Model 2 by looking at the appropriate variance parameter in the second-stage covariance matrix, Σ, or by looking at histograms of posterior means of the β_i. This will describe the extent and form of the heterogeneity suggested by the data given this particular model specification. However, this does not provide a formal test for the existence of heterogeneity in returns to education, and to this end, we consider the following specification:

Model 3: $\alpha_i | \alpha, \sigma^2_\alpha \overset{iid}{\sim} N(\alpha, \sigma^2_\alpha), \;\; \beta_i = \beta$.

In Model 3, we permit heterogeneity in intercepts only and restrict everyone to have a common return to education. By comparing Model 3 against Model 2, we are able to test if the data favor a model which *additionally* permits heterogeneity in returns to schooling to one which 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 this second-stage specification is its 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.

[12] 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, Levy and Willett (1995), Grogger and Eide (1995), Heckman and Vytlacil (2001), and DiNardo and Tobias (2001).

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 could be approximated quite accurately with a discrete distribution with a suitable number of support points. This leads to our fourth model, involving heterogeneity of a discrete nature.

Model 4: $\theta_i = \theta_g^0$ with probability π_g for $g = 1, 2, \cdots G$.

In this model, 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. We compare these models by computing marginal likelihoods associated with a variety of different G, and also compare each of these models to the continuous alternatives in Models 2 and 3.

It is interesting to note that the assumption of discrete heterogeneity has a precedent in this literature. For example, Ichino and Winter-Ebmer (1999) suppose that there are exactly four different “types” of individuals defined as those having either a low or high marginal cost of and marginal return to education. Although their estimation approach made use of various instrumental variables, the type of heterogeneity they assume can be represented exactly by discrete distribution described above. The benefit of our estimation approach is that we are able to test if heterogeneity exists, and can also test if a discrete distribution provides the best way to characterize it.

In our empirical work, we also generalize Model 4 by adding covariates to help explain the mapping of individuals to the various subgroups:

Model 4W $\theta_i = \theta_g^0$ with probability $\pi_{gi} = \pi_g(w_i), \;\; g = 1, 2, \cdots G$.

In this model, we introduce time-invariant observables (w_i) to help explain the assignment of intercept and slope parameters to various individuals in the sample. This feature, as shown in the following section, will prove to be very useful for the evaluation of hypothetical interventions. Intuitively, in this model individuals are sorted into G different groups. The probability of individual i being in a given group depends on observable characteristics (w_i) such as measured cognitive ability. The idea is that people of the highest ability may be most likely to receive the highest return to schooling, while people of slightly lower ability may be most likely to receive a slightly lower return. We thus use an ordered probit specification to model this probability. Such a specification can be interpreted as implying that a Normally distributed latent variable drives the allocation into groups.[13]

To be consistent with the latent variable interpretation, we also impose the following ordering of the slope coefficients in the components of the mixture:[14]

$$\beta_1^0 < \beta_2^0 < \cdots < \beta_G^0.$$

As mentioned previously, the reason for imposing this ordering is that we might think the observables w are related monotonically to the return to schooling parameters β_g^0. That is, if the latent index is large enough, then the individual is most likely to be drawn from the highest return component, while if the latent value is slightly lower, then she is most likely to be drawn from the second highest component, *etc.*

Given the latent variable interpretation and the ordering, our model amounts to a (restricted) finite mixture model for log wages coupled with an ordered probit to explain the component assignment mechanism. Since this model is less-widely used than the models discussed previously, it is useful to briefly describe our procedure for estimating it, though complete discussion is provided in Appendix B.

We fit this model by first augmenting the parameter space with a set of component indicators, say $\{c_{gi}\}$, $g = 1, 2, \cdots G$, where $c_{gi} = 1$ indicates that the i^{th} individual is drawn from the g^{th} component of the mixture. As stated before, an alternative way of interpreting this is in terms of a latent variable c_i^*, where

$$c_i^* | w_i, \zeta \overset{ind}{\sim} N(w_i \zeta, 1).$$

In this setting, we observe $c_{gi} = 1$ iff $\tau_{g-1} < c_i^* \leq \tau_g$, with $\tau_0 = -\infty$ and $\tau_G = \infty$. Hence,

$$\Pr(c_{gi} = 1 | w_i, \tau_g, \tau_{g-1}, \zeta) = \Phi(\tau_g - w_i \zeta) - \Phi(\tau_{g-1} - w_i \zeta),$$

where $\Phi(a)$ denotes the c.d.f. of the standard Normal distribution evaluated at point a.

A complete description of the complete conditionals for this model and priors employed for the second stage parameters are provided in Appendices B and C.

[13] The traditional derivation of the ordered probit model in a qualitative choice context involves a random utility model. In this case, the latent variable driving the choice problem is utility. Our context is, of course, different here, but the idea of utility driving the individual as "choosing" to be in group g provides useful intuition.

[14] As a practical matter, this restriction can be incorporated formally into the prior. As an alternative, we might place different informative priors over each β_g to approximately incorporate the restriction. The latter approach offers a compuatationally-attractive alternative.

Finally, to further investigate the appropriateness of the second-stage Normality assumption, we also consider a generalized family of models:

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

In the above, we let $f\left(\theta_i|\cdot\right)$ denote the p.d.f. of θ_i and assume independence over i. Furthermore, we let $\phi(x; \mu, \Sigma)$ denote the multivariate Normal density for x with mean μ and covariance matrix Σ. Thus, in Model 5, we assume that the heterogeneity distribution follows a finite mixture of Normals. This offers a flexible modeling alternative that is capable of capturing a variety of possible forms for the heterogeneity. Indeed, mixtures of Normals are so flexible that methods based on them are often interpreted as being analogous to classical nonparametric methods. Like Model 4, the above model assumes the population is comprised of a discrete number of different groups, but unlike Model 4, this model permits individuals within each group (or component of the mixture) to possess different intercepts and returns to education. By computing marginal likelihoods associated with Model 5 for various values of $G \geq 2$, we are able to test for the appropriateness of the second-stage Normality assumption, and also are able to test if these generalized continuous heterogeneity models are preferred over discrete alternatives.

3.1 A Note on Policy Evaluation

To this point we have introduced a set of competing models for describing the nature of individual-level heterogeneity, tied these models to forms employed in previous work, and described procedures for selecting the specifications most favored by the data. In this section, we move beyond this and describe how one can use the results obtained to predict distributions of returns to education for various subpopulations when some policy intervention is introduced. Specifically, we focus our attention on providing answers to two questions which have received attention in this literature: (1) How can we learn about the characteristics of individuals who are affected by or *comply* with a given intervention (*e.g.* attend college as a result of a tuition subsidy)?[15] (2) How might we use this updated information to characterize the distribution of returns to schooling for those complying with a given intervention? Our efforts in this regard are quite similar in spirit to the "treatment effects" literature, and the parameter we work with is similar to a Local Average Treatment Effect (LATE) parameter (Imbens and Angrist (1994)) which measures expected returns to schooling for the subgroup affected by a particular

[15] This issue was addressed in Kling (2001) who finds that individuals "complying" with a distance-to-college instrument are more likely to come from disadvantaged backgrounds, and thus this IV estimate captures a return which is more heavily-weighted toward these types of individuals.

instrument. Not only will our analysis enable us to obtain point estimates of such mean effects, but will also enable us to recover the *entire posterior predictive distributions* of returns to schooling for various subpopulations. Further, such distributions are not limited to LATE-type analysis, but can be extended to recover distributions of returns for a variety of different subpopulations.

To fix ideas on a particular parameter of interest, let us suppose that the government is considering implementing a tuition subsidy to encourage future investments in schooling. Assume that, provided an individual attends college, her "return" is measured as her marginal benefit from an added year of education. We begin by assuming that the parameter of interest is the return to schooling *for those who actually choose to attain an additional year of schooling as a result of the tuition subsidy.*

Before going into formal details, we first provide some intuition. Knowing that an individual has changed schooling decisions because of the existence of the intervention leads us to update our beliefs about the distribution of her explanatory variables. In other words, *the decision to comply itself leads us to update our beliefs about the characteristics of the "compliers."* In terms of our example, if a given person can be induced to attend college because of a subsidy, then it is unlikely that the person will be of, say, very high or very low ability. If the person were of very high ability, then she would have been likely to attend college regardless of the existence of the subsidy, while if the person was of low ability, then she would be unlikely to enter college even with the subsidy. Since explanatory variables like measured cognitive ability might also be empirically important explanatory variables in the heterogeneity distribution, (*e.g.* in Model 2W or 4W), the distribution of returns for those who "comply" with the subsidy can be quite different from the population distribution of returns.

Since we are thinking of hypothetical interventions that may affect some future population, let β_f denote the future return to schooling of this population - the percent change in log wages resulting from completing an additional year of education. We continue to assume that the model described in (4) - (8) correctly describes log wages, as well as the heterogeneity of returns for this population.

To establish the formal results, we introduce an auxiliary model to describe an individual's decision to attain an additional year of schooling, and think of this as the decision to attain a year of college education. We suppose that this decision depends upon a set of covariates w_f as well as a tuition variable T. The variables in w_f include things like family characteristics and measured ability (*i.e.* variables also assumed to affect our heterogeneity distribution in Models 2W and 4W).

We assume a latent index model for the college entry decision with link function F which depends on

w_f, tuition (denoted as T), and parameters Γ (*e.g.* $F(w_f, T; \Gamma) = \Phi(w_f\gamma_1 + T\gamma_2)$). In this case, the probability that an individual attends college at a lower (subsidized) tuition level T_0', but does not attend at the previous tuition level T_0 is given as

$$\Pr\big[D_f(w_f, T_0'; \Gamma) = 1, D_f(w_f, T_0; \Gamma) = 0 | w_f, T_0', T_0, \Gamma\big] = F(w_f, T_0'; \Gamma) - F(w_f, T_0; \Gamma),$$

where $D_f(w_f, T; \Gamma)$ is the binary variable which equals one if the individual takes an additional year of schooling, and equals zero if the individual does not attain additional schooling. For the sake of notational convenience, we will simply denote these decisions as $D_f' \equiv D_f(w_f, T_0'; \Gamma)$ and $D_f \equiv D_f(w_f, T_0; \Gamma)$, leaving the dependence on w_f, T and Γ implicit. We regard T_0 and T_0' as constants, and note that different values of T_0 and T_0' simply define different functions of interest.

Now, let us revisit our original questions of interest. First, we wanted to learn about the characteristics of individuals who are affected by or *comply* with a given intervention. Second, we wanted to use this information to predict the distribution of returns for the population that would attain additional schooling as a result of the tuition subsidy. Thus, our final goal would be to derive:

$$p(\beta_f | D_f' = 1, D_f = 0, T_0, T_0', \text{Data}),$$

the posterior predictive return distribution for those not attending college at T_0, but attend at T_0'. With a bit of work, it becomes clear how one might proceed to calculate this predictive:

$$\begin{aligned} p(\beta_f | D_f' = 1, D_f = 0, T_0, T_0', \text{Data}) &= \int_{R_{w_f,\Gamma,\lambda}} \big[p(\beta_f | w_f, \Gamma, \lambda, D_f' = 1, D_f = 0, T_0, T_0', \text{Data}) \cdot \qquad (9) \\ &\quad p(w_f, \Gamma, \lambda | D_f' = 1, D_f = 0, T_0, T_0', \text{Data})\big]\, dw_f d\Gamma d\lambda \\ &= \int_{R_{w_f,\Gamma,\lambda}} \big[p(\beta_f | w_f, \lambda) p(w_f, \Gamma | D_f' = 1, D_f = 0, T_0', T_0, \text{Data}) \qquad (10) \\ &\quad p(\lambda | \text{data})\big]\, dw_f d\Gamma d\lambda, \end{aligned}$$

where the integration is taken over the support of w_f, Γ and λ (denoted $R_{w_f,\Gamma,\lambda}$). Remember that λ was our general notation for the parameters in the hierarchical prior in our models. The second line follows since the return distribution is assumed to depend only on the characteristics w_f and parameters λ as in equations (4) - (8),[16] and the parameters of the heterogeneity distribution (λ) only depend on the past data. Consistent with our reduced-form view, we are also assuming that the parameters of the binary choice model are independent of the parameters contained in the heterogeneity distribution.

The conditional distribution $p(w_f, \Gamma | D_f' = 1, D_f = 0, T_0', T_0, \text{Data})$ can be simplified and calculated as

[16] This assumption implies, for example, that unobservables making it more likely for an individual to attain further education are uncorrelated with the return (β_f) from the receipt of that education. If this assumption were not true, then we would have to carry along the conditioning information $D_f' = 1$ and $D_f = 0$ in the conditional for β_f in (10). See Heckman and Vytlacil (1998) for further discussion of such a model.

follows:

$$\begin{aligned} p(w_f, \Gamma | T_0', T_0, D_f' = 1, D_f = 0, \text{Data}) &= \frac{\Pr(D_f' = 1, D_f = 0 | T_0', T_0, w_f, \Gamma, \text{Data}) p(\Gamma|\text{Data}) p(w_f)}{\Pr(D_f' = 1, D_f = 0 | T_0', T_0, \text{Data})} & (11) \\ &\propto [F(w_f, T_0'; \Gamma) - F(w_f, T_0; \Gamma)] p(\Gamma|\text{Data}) p(w_f), & (12) \end{aligned}$$

where in (11) and (12) we use the fact that the future covariates w_f are independent of the binary choice parameters Γ. Also note from (12) that *the joint posterior distribution of characteristics and parameters for those who comply is different from the (unconditional) joint posterior distribution of characteristics and parameters.* Thus, the derived expression in (12) provides a vehicle for updating our beliefs about characteristics of the compliers, and thus enables us to address our first question. For estimation purposes will also assume (arguably quite reasonably) that the distribution of future characteristics is the same as the population distribution of characteristics (*i.e.* $p(w_f) = p(w)$). We put the result in (12) together with our previous expression for the posterior predictive return distribution in (10) to obtain:

$$\begin{aligned} p(\beta_f | D_f' = 1, D_f = 0, T_0, T_0', \text{Data}) &\propto \int_{R_{w_f, \Gamma, \lambda}} p(\beta_f | w_f, \lambda) \left[F(w_f, T_0'; \Gamma) - F(w_f, T_0; \Gamma) \right] \cdot & (13) \\ & \quad p(\Gamma|\text{Data}) p(\lambda|\text{Data}) p(w_f) dw_f d\Gamma d\lambda. \end{aligned}$$

Straight-forward arguments also show that it is also possible to obtain the (average) distribution of returns in the population:

$$p(\beta_f|\text{Data}) = \int_{R_{w_f, \lambda}} p(\beta_f | w_f, \lambda) p(\lambda|\text{Data}) p(w_f) dw_f d\lambda, \tag{14}$$

or the distribution of returns for those attending college at a given value of tuition:

$$p(\beta_f | T_0', D_f' = 1, \text{Data}) \propto \int_{R_{w_f, \Gamma, \lambda}} p(\beta | w_f, \lambda) [F(w_f, T_0'; \Gamma)] p(\Gamma|\text{Data}) p(\lambda|\text{Data}) p(w_f) dw_f d\Gamma d\lambda. \tag{15}$$

Several observations are in order regarding these predictives. First, note that the predictives in (13) - (15) can be different from one another, as the conditioning information contained in these predictives leads us to marginalize over different distributions. Second, note that if the covariates w_f have no role in the heterogeneity distribution, then all of the predictives reduce to the same distribution. In this case, we may update our beliefs about the characteristics of the compliers, but this updating is irrelevant since these characteristics play no role in the distribution of returns. Finally, note that means of these distributions are similar in spirit to various "treatment parameters" which have received substantial attention in the program evaluation literature. Specifically, the mean of (13) is similar to the Local Average Treatment Effect (LATE), the mean of (14) the Average Treatment effect (ATE), and the mean of (15) the effect of Treatment on the Treated (TT) (*e.g.* Heckman and Robb (1985)), where the "outcome" is regarded as the marginal return to education. Finally, note that the method described here provides the *entire posterior distributions of returns for the various subgroups*, so that means, variances, percentiles, etc. can be computed in a straightforward manner.

In order to compute these predictives, as suggested in (13), we need to carry out a relatively high-dimensional integration problem. However, calculation of these predictives is relatively straightforward via a simulation-based approach. Since β_f is a scalar, we could simply fix various values β_f^0 and estimate the ordinates of the predictive at these different values via Rao-Blackwellization. That is, for each draw from the posterior distribution of Γ and λ as well as the draw from $p(w_f)$, we compute the value of the integrand, do this for all draws, and take a sample average of the result to obtain a simulation-consistent estimate of the ordinate. Repeating this for all values of β_f^0 gives us a discretized (and as yet unnormalized) estimate of the desired posterior.

The biggest complication, of course, arises in the estimation of $p(w_f)$. For small-dimension w_f, one possibility is to estimate its distribution nonparametrically, and use this nonparametric estimate to compute the desired integrals, as in (14). In this case, to estimate the ordinate in (13) we could calculate:

$$p(\beta_f^0|D_f' = 1, D_f = 0, T_0, T_0', \text{Data}) = \int_{R_{w_f}} \left(\frac{1}{M} \sum_{i=1}^{M} p(\beta_f^0|w_f, \lambda^{(i)}) \left[F(w_f, T_0'; \Gamma^{(i)}) - F(w_f, T_0; \Gamma^{(i)}) \right] \right) \hat{p}(w_f) dw_f \quad (16)$$

using Simpson's rule or a trapezoidal rule to do the integration over w_f given our nonparametric estimate $\hat{p}(w_f)$.[17] We also have made use of the notation $\lambda^{(i)}$ and $\Gamma^{(i)}$ to denote draws from the posterior distribution of these variables. Repeating this process for a variety of different β_f^0 and normalizing the result gives an estimate of the desired predictive. For higher dimension problems, discrete approximations, or perhaps even parametric assumptions regarding the distribution of w_f could be used to facilitate computation.

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 U.S. ranging in age from 14-22 as of the first interview date in 1979. Importantly for our purposes, the NLSY contains detailed information on the earnings and wages, educational attainment, family characteristics, and test scores of the sampled individuals. For this application, we use a version of the NLSY which allows us to obtain an earnings

[17] Alternatively, one might approximate $p(w_f)$ using a discrete distribution, and draw from this distribution as well to calculate the desired integrals via simulation.

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 labor 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 dollars 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. As such, we are careful to delete individuals whose education is clearly mis-measured.

Given this sample selection scheme, we obtain data on $N = 2,269$ individuals for a total (denoted NT) of $NT = 18,543$ person-year observations. Our time-varying characteristics include potential labor market experience and its square[18], a time trend coded so that 1979 values take on the value of 1 and increments by 1 for consecutive years, and a continuous measure of the local unemployment rate in the given year. Our set of time-invariant characteristics include a standardized test score,[19] 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.

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 individuals who are quite active in the labor 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

[18] Potential experience is defined as Age - Education - 5.

[19] 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).

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.

To investigate this issue rather broadly, we consider the wage outcomes of white males aged 16-18 in 1981.[20] 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 of the appendix 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. It is also interesting to note that the 1993 densities tell a quite different, and certainly expected story - the wage distribution for those with at least a high school degree is clearly shifted to the right relative to those without this degree.

5 Empirical Results

The set of competing models outlined in the previous section are estimated using the Gibbs sampler described in Appendix B. In our Gibbs runs, we take 11,000 replications. We discard the first 1,000 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

[20] The use of 1981 data is essentially without loss of generality - similar results hold for the other early years in the NLSY. However, we can not obtain wage data for similarly aged individuals at later years of the sample, since the same individuals are tracked over time and age at each consecutive interview year.

runs, and the use of *blocking* or *grouping* steps also helps to mitigate autocorrelation in the chains. The priors are described in Appendix C. Suffice it to note here that we center 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 (in a sense explained in Appendix C). 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. To investigate these questions, we calculate log marginal likelihoods[21] associated with Models 1, 2, 3, 4 and 5 using the method proposed by Chib (1995) (see Appendix B for details). For Model 4, we report results using $G = 10$. We found, after extensive search through a variety of Model 4's with varying numbers of support points G, that the value $G = 10$ was most supported by the data. This value produced the highest marginal likelihood among the competing Model 4's, and the maximized likelihoods were not found to significantly increase after adding more support points. For Model 5, we find $G = 1$ to be the optimal number of components, though this model is identical to the bivariate Normal heterogeneity of Model 2. Given this, we present Model 5 results in Table 1 using $G = 2$.

To put all models on an equal footing, we do not, as yet, include observable variables w in any of the second-stage equations. To illustrate that our results are not sensitive to the choice of prior, we also calculated the Schwarz or Bayesian Information Criterion (BIC) using our non-informative prior and present this in Table 1.

Table 1: Model Comparison Results

	Model 1: $\alpha_i = \alpha$, $\beta_i = \beta$	Model 2: $\theta_i \sim N(\theta_0, \Sigma)$	Model 3: $\alpha_i \sim N(\alpha, \sigma_\alpha^2)$, $\beta_i = \beta$	Model 4: $\theta_i = \theta_g^0$ with prob π_g, $G = 10$	Model 5: Normal mix. $G = 2$
Log Marg. Likelihood	-12,859	-8,399	-8,523	-8,621	-8,456
BIC	-25,297	-16,577	-16,747	-17,077	-16,608

[21] The ratio of two marginal likelihoods is the Bayes factor comparing the two relevant models. If equal prior weight is attached to each model, the Bayes factor is equivalent to the posterior odds ratio.

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 results in Table 1 are overwhelming. The continuous, Normally distributed heterogeneity of Model 2 is massively preferred by the data. The second preferred choice is Model 5, which mixes two Normal distributions, followed by Model 3 which permits Normal heterogeneity in intercepts only and restricts returns to schooling to be constant across individuals. By comparing Model 3 to Model 2, *we learn that the data favor a model that additionally permits returns to schooling to vary across individuals to one which allows heterogeneity across individuals to be captured only through baseline differences in intercepts.* This provides strong evidence that returns to schooling are heterogeneous, and in terms of our underlying structural model, this provides some evidence of heterogeneity in both marginal costs of and marginal returns to schooling across individuals. The discrete heterogeneity model ranks fourth among our five competing models, and seems to provide a clearly inferior description of the heterogeneity.[22] Though the data have shown that it is important to permit individuals to differ in intercepts and slopes, the data also rejects a model that characterizes these differences by supposing that the population consists of a finite number of different groups, with "identical" individuals within each group. 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)). We also stress that these strong findings and the ranking of the models was not affected by our choice of prior, as shown in the last row of Table 1 which provides results using a non-informative prior.

In Table 2 below, we present results of key first and second-stage coefficients from Models 1,2, and 4, and will refer to these results in the discussion to follow. For Model 4, we present results for the $G = 2$ case. Note that this 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.

[22] It is interesting to note that Allenby and Rossi (1999) also found a strong preference for a continuous distribution in the second stage over discrete alternatives in a marketing application.

Table 2: 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	.106	.121	.111	
	(.004)	(.004)	(.004)	
Experience^2	-.003	-.003	-.003	
	(.0002)	(.0001)	(.0001)	
Time	-.023	-.025	-.019	
	(.002)	(.003)	(.003)	
Unemp. Rate	-.010	-.004	-.005	
	(.001)	(.001)	(.001)	
Second-Stage Parameters			1^{st} Comp. (θ_1^0)	2^{nd} Comp. (θ_2^0)
Intercept	.595	.300	.387	.679
	(.032)	(.058)	(.044)	(.046)
Education	.105	.116	.090	.112
	(.002)	(.005)	(.003)	(.003)

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 after controlling for other covariates, real log wages tended to decline by about 2 percent each year. We interpret this result, however, with caution as the time trend is highly correlated with our potential experience measure (with a correlation coefficient equal to .75). Expected wages are clearly increasing for an individual in their next year of the panel, as the potential experience effect washes out the negative effect of the time trend.[23]

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 Table 1. 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,

[23] 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.

the posterior means of the elements of Σ are: $E(\Sigma_{11}|Data) = 0.8341$, $E(\Sigma_{12}|Data) = -0.065$ and $E(\Sigma_{22}|Data) = 0.0058$, 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 0.162 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 2 plus remember that $E(\Sigma_{22}|Data) = 0.0058$) imply β_i (the individual-level return to schooling parameter) is roughly $N(.116, .0058)$. This implies a 95% probability interval would be roughly $[-.03, .27]$, indicating that the return to an added year of schooling varies from minus three percent per year to twenty seven percent 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 Table 1. 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 [-1.49, 2.09].

It is also very interesting to note that this 95% probability interval for returns to schooling "covers" the majority of point estimates reported in previous studies, and thus our framework might be capable of reconciling the disparity in previous results. For example, Card (2001, Table II pages 1146-1148) reviews eleven studies using supply-side instrumental variables and reports the point estimates associated with the use of both IV and OLS in each study. With the exception of the study by Ichino and Winter-Ebmer (1999), which uses German data and does not provide a comparable returns to schooling measure (i.e. they measure schooling as a dummy for completion of at least high school), *all of these point estimates fall within our 95% probability interval.* The range of point estimates in previous studies is both consistent with the results suggested by our study as well as the interpretation of the IV estimand when returns are heterogeneous. Since IV can be interpreted as recovering a weighted average of returns for different subgroups in the population, it is to be expected that different instruments will yield different point estimates, and these estimates should fall within our probability interval which directly measures the extent of unobserved heterogeneity in returns.[24]

Thus far, we have shown that the second-stage normality assumption provides a better description of the bivariate heterogeneity than any of the remaining competitors in Models 1-4. As shown in Table 2, the "closeness" of the β_g^0 for each of the two components together with the fact that the

[24] It is also important to recognize that the eleven studies reviewed use very different sources of data (as reported in Card (2001)).

associated component probabilities are virtually identical (.51 for the component with return .09, and .49 for the component with return .112) also provides suggestive evidence of the appropriateness of the Normality assumption in the second stage. To provide some further insight on this issue, and to see if our normality assumption is a reasonable one, we plot histograms of the posterior means of the intercepts and slopes for for every individual in the sample (*i.e.* $E(\alpha_i|Data)$ and $E(\beta_i|Data)$ for $i = 1, .., N$) using Model 2. These are presented in Figures 2 and 3 of the appendix, and a quick inspection of these tables suggests that our normality assumption does not seem to be greatly at odds with the data. Of course, a discrete distribution with enough points of support can always be used to approximate a continuous distribution. It appears that this is what results using Model 4 reveal. In particular, with Model 4, the data choose $G = 10$ points of support as being optimal. A plot of these 10 points of support against the probability associated with each reveals that the discrete distribution of Model 4 approximates the histograms in Figures 2 and 3 quite well.[25] However, since Model 4 with $G = 10$ contains many more parameters than the relatively parsimonious Model 2, the latter is strongly supported using standard model comparison methods.

We can investigate the appropriateness of the Normality assumption more formally by estimating Model 5 which introduces a finite mixture of Normals to the second-stage. The results of Table 1 indicate that a single Normal distribution is the preferred choice for the heterogeneneity distribution. That is, mixing two Normals together (i.e. $G = 2$) causes only a slight improvement in fit relative to a single Normal distribution. This improvement in fit is outweighed by the reward for parsimony in model comparison measures such BIC and the marginal likelihood. Since empirical results (*e.g.* point estimates of parameters) are very similar for Model 5 with $G = 2$ and $G = 1$ we do not provide additional empirical results for Model 5 here.

In summary, we have found strong evidence in favor of Normally distributed heterogeneity in returns to schooling. This evidence is supported by examination of point estimates obtained from the alternate models as well as the results obtained from more formal model comparison methods. 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.

[25] Detlails of these calculations are available upon request.

5.3 *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 our model in attempt to explain differences in returns to schooling across individuals. To this end, we first consider results obtained from Model 2W. Remember that this takes Model 2 (our most preferred bivariate Normal heterogeneity model), and adds to it a measure of cognitive ability, highest grade completed by the respondent's mother and father, an indicator for residence in a broken home at age 14, and number of siblings as explanatory variables in both the intercept and slope equations. These ability and family background variables may help to explain baseline differences in earnings across individuals, and may also help to explain why returns to schooling differ across individuals.

Table 3 contains posterior means and standard deviations associated with these second-stage parameters. Also provided in Table 3 are Bayes factors in favor of each coefficient being equal to zero. A common rule of thumb (see, e.g., Poirier, 1995, page 380) uses Bayes factors of less than .10 to indicate strong evidence (and values between .10 and 1.0 to indicate slight evidence) against the hypothesis that the relevant coefficient equals zero.

While no obvious story emerges from Table 3, we do see that ability and family size (Numsibs) have a reasonably strong effect through the intercept, but little evidence that marginal returns to schooling are affected by either Ability or Numsibs. Other variables such as Mother's education (Momed), which seem to be "significant" (and negative!) in the intercept equation are often counteracted through the interaction with education. For example, at 12 years of schooling, changes in Momed have very little impact on the wage equation, while the effect of Momed on wages increases as the child attains more schooling. We do find some weak evidence that reporting being from a broken home (Broken) has an effect both through the intercept and slope and that being from a broken home has a larger negative effect on earnings for those that attain more schooling. Seemingly the most significant results to take away from this table are that some variables, such as measured cognitive ability and number of siblings play a strong role in explaining heterogeneity across intercepts, yet no variables seem to play a major role in explaining variation in returns to schooling across individuals.

Table 3: 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	1.028 (0.222)	$--$	0.058 (0.018)	$--$
Ability	0.065 (0.056)	0.047	0.002 (0.005)	24.191
Momed	-0.025 (0.022)	0.106	0.002 (0.002)	1.740
Daded	-0.010 (0.017)	3.466	0.001 (0.001)	26.505
Broken	0.0853 (0.121)	0.373	-0.011 (0.010)	0.644
Numsibs	-0.0313 (0.023)	0.027	0.002 (0.002)	0.909

To see this last point more clearly, and to investigate how much these explanatory variables "matter," 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. Remember that, with Model 2, we found $E(\Sigma_{11}|Data) = 0.8341$ and $E(\Sigma_{22}|Data) = 0.0058$. With Model 2W, we have variables which we would expect should explain away some of the heterogeneity and, thus, these point estimates should be smaller. In fact, for Model 2W we find the point estimates of these variances to be only slightly smaller with $E(\Sigma_{11}|Data) = 0.8163$ and $E(\Sigma_{22}|Data) = 0.0056$. In other words, the addition of these time-invariant characteristics in Model 2W explained away only 2.1 percent of the second-stage intercept variation, and only 3.6 percent of the second-stage variation in returns to schooling.[26] *Thus, we see very little reduction in the total amount of variation across individuals after including these time-invariant characteristics, and thus conclude that the role of unobserved heterogeneity remains substantial.*

5.4 *A Note on the Evaluation of Policy*

In this section, we apply the general methodology outlined in section 3.1 to calculate the (posterior predictive) return to schooling distributions for various subpopulations. To this end, we suppose the government is considering implementing a tuition subsidy, along the lines of the framework outlined in section 3.1. We imagine that the subsidy may affect the decision to enter college of a subgroup of high school graduates, and thus seek to characterize the distribution of returns for this group of

[26] We do not report these results for Model 4W or 5W as substantive conclusions are not affected.

"compliers." We assume that the "return" for the individuals who do enter college as a result of the subsidy is measured as the their marginal return to an added year of schooling (β_f). In addition to obtaining this posterior predictive distribution of returns for those who attend after the subsidy, but did not attend before (as in (9)), we will also recover the posterior predictive distribution of returns in the population (as in equation 15), and the posterior predictive distribution of returns for those who attend college at a given level of tuition (as in (16)).

To make things computationally simple, we limit the dimension of characteristics in w_f and thus model our college entry decision as a function of state-level tuition and a measure of cognitive ability.[27] The tuition variable used is the average tuition at 4-year colleges within the respondents' state of residence at age 18, adjusted to real 1986 dollars. We then re-run Model 2W to include only a constant and our measure of cognitive ability as explanatory variables in the second-stage heterogeneity distribution.

We assume a probit model[28] applies to the decision to attain at least one year of schooling for a sample of high school graduates. We thus take data on 1,775 white males in the NLSY who attain at least 12 years of education (as of 1993) and model their decision to get at least one year of college education as a function of an intercept, state-level tuition (measured in hundreds of 1986 dollars) and our measure of cognitive ability. We present in Table 4 below posterior means and standard deviations of coefficients and marginal effects from this simple probit model:

Table 4:
Posterior Means and (Standard Deviations)
of Coefficients and Marginal Effects
from Probit Model

	Coeff	Marg. Effect
Constant	.280	
	(.106)	
Ability	.662	.264
	(.037)	(.015)
Tuition	-.015	-.006
	(.006)	(.002)

As expected, our ability measure is a very strong predictor of the decision to attain at least one more year of schooling. The state level tuition variable has a negative effect on the likelihood of college entry, and its posterior places very little mass over zero. The size of the effect, however, is rather small

[27] We gratefully acknowledge the support of Jim Heckman and Jingjing Hsee in acquiring this tuition data.

[28] Since MCMC methods for probit models are well-known (see Albert and Chib, 1993), we do not provide details here. We use a relatively noninformative but proper prior, setting prior means of the probit coefficients equal to zero and the prior covariance matrix equal to 4 I_3, with I_n denoting the $n \times n$ identity matrix.

as a \$100 reduction in state-level tuition leads to less than a 1 percent increase in the likelihood of attaining one more year of post-secondary education.

In Table 5 below, we use this information, as described in section 3.1, to determine the posterior predictive return to schooling distributions for various subpopulations. As discussed in that section, the event that individuals acquire more schooling at the given level of tuition, or acquire more schooling at the lower tuition level but would not acquire more schooling at the pre-subsidy level leads us to update our beliefs about the distribution of her explanatory variables. Since the return distribution may also depend on these explanatory variables, the resulting distribution of returns for the "compliers" may be quite different than the population distribution of returns. For this application, however, the small role of ability in the second-stage return to schooling distribution[29] implies that our predictive distributions in (13)-(15) for the various subgroups should be extremely similar, as pointed out in section 3.1.

To show that the distributions of characteristics differ for these subpopulations, and thus the distributions of returns may also differ, we present in Figure 4 below the result of one particular experiment. In this experiment, we nonparametrically estimate the population distribution of ability, and then, after fixing the binary choice parameters to their posterior means, plot the distribution of ability for those who would be induced to attend college after lowering tuition by one standard deviation ($T_0' = \overline{T_0} -$ \$544), but would not have attended when state level tuition was at its overall mean value ($T_0 = \overline{T_0}$ = \$1,700).[30] As shown in this table, these densities are slightly different, as the tuition variable was found to play a small role in the binary choice model.[31] *These results do, however, show that the "compliers" would be less likely to have either very low or very high values of ability, as low-ability individuals would be very difficult to induce to attend college at T_0', and high-ability individuals would be very likely to attend at the lower tuition level $\overline{T_0}$.*

We then use the results suggested by Figure 4 to estimate features of the posterior predictive return distributions for various subpopulations. In all cases, we use results obtained from the Normal heterogeneity distribution of Model 2W, given our previous evidence that bivariate Normality provides an adequate description of the heterogeneity. These results are presented in Table 5 below.

[29] Recall from Table 4 that ability played a strong role in the intercept distribution, but a small role in the return to schooling distribution. Similar results are obtained here when dding ability as the only explanatory variable in the second-stage heterogeneity distribution.

[30] See equation (12) for a derivation.

[31] Also note that this difference is not attributable to any kind of estimation error, since we base results of the "compliers" density on the estimates of $p(w)$.

Table 5:
Posterior Quantities Associated with Various
Predictive Return Distributions

	Post. Mean	Post. Std. Dev	Pr$(\cdot > 0 \vert \text{Data})$	Pr$(\cdot > .10 \vert \text{Data})$
$\beta_f \vert \text{Data}$	.096	.073	.90	.45
$\beta_f \vert D'_f = 1, T'_0, \text{Data}$	.097	.073	.90	.46
$\beta_f \vert D'_f = 1, D_f = 0, T'_0, T_0, \text{Data}$	.096	.073	.90	.45

As expected, the results for the various subgroups are virtually identical since the ability variable was found to play a very small role in explaining variation in our return to schooling distribution. In this sense, we have updated our beliefs about the characteristics of individuals in each subgroup (as shown in Figure 4), though this updating has no effect on the resulting predictive return to schooling distributions. Hence, we find for this application that the distribution of returns for an "average" person, for the "treated" or for the "compliers" are essentially identical - a useful result to know when evaluating educational policy. We stress, however, that our methods are quite operational, and thus offer a seemingly promising alternative for calculating returns to "treatment" for different subgroups when returns to that treatment are heterogeneous.

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.

We also showed how use of our methods provides the quantities needed for evaluating the effects of hypothetical policy interventions, including parameters such as the Local Average Treatment Effect (LATE) and the effect of treatment on the treated (TT). In addition, we argued that the disparity of IV estimates in previous work is perfectly consistent with, and perhaps rationalized by, the extent of heterogeneity we find in this investigation.

The models used in this paper are flexible enough to establish the main empirical patterns in our data set (*i.e.* heterogeneity does exist and it is better modeled by a continuous than a discrete distribution). However, it is worth noting that our models generalize quite easily since they all involve mixtures of Normal distributions. A great deal of the recent Bayesian statistical literature has focussed on developing tools for such mixtures. For instance, Carlin and Polson (1991), and Geweke (1993) develop methods for a certain type of scale mixture of Normals which is equivalent to the Student-t distribution. Dirichlet mixtures of Normals (see, e.g., Escobar and West (1996)) can be used if the researcher wishes to approximate f using methods of comparable flexibility to nonparametric kernel smoothing. Thus, the general framework discussed in this paper is a very powerful one that can be extended in many ways which might be of interest for a variety of other applications.

7 Appendix A: Figures

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.

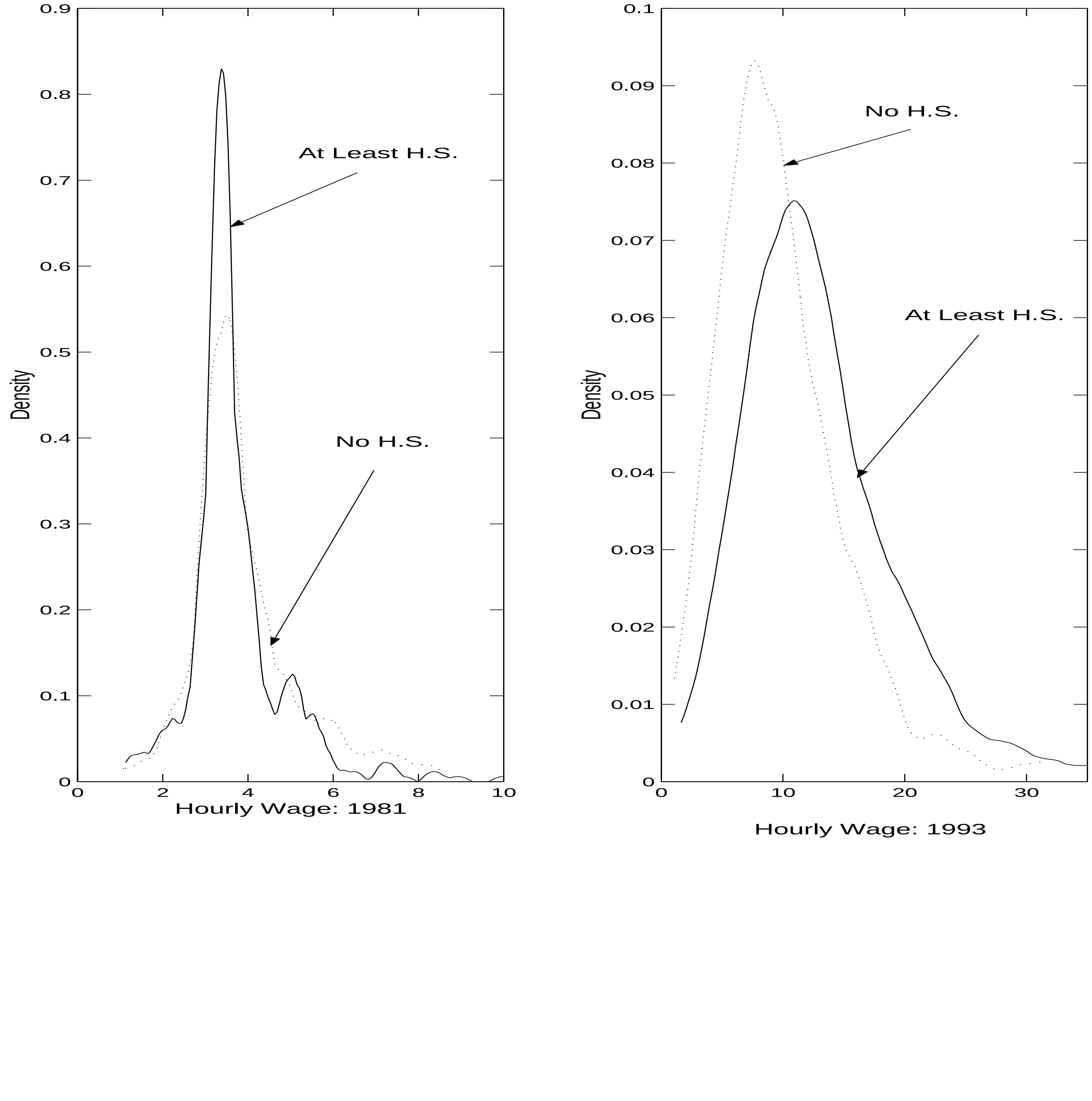


Figure 2: Histogram of Posterior Means of α_is for Model 2

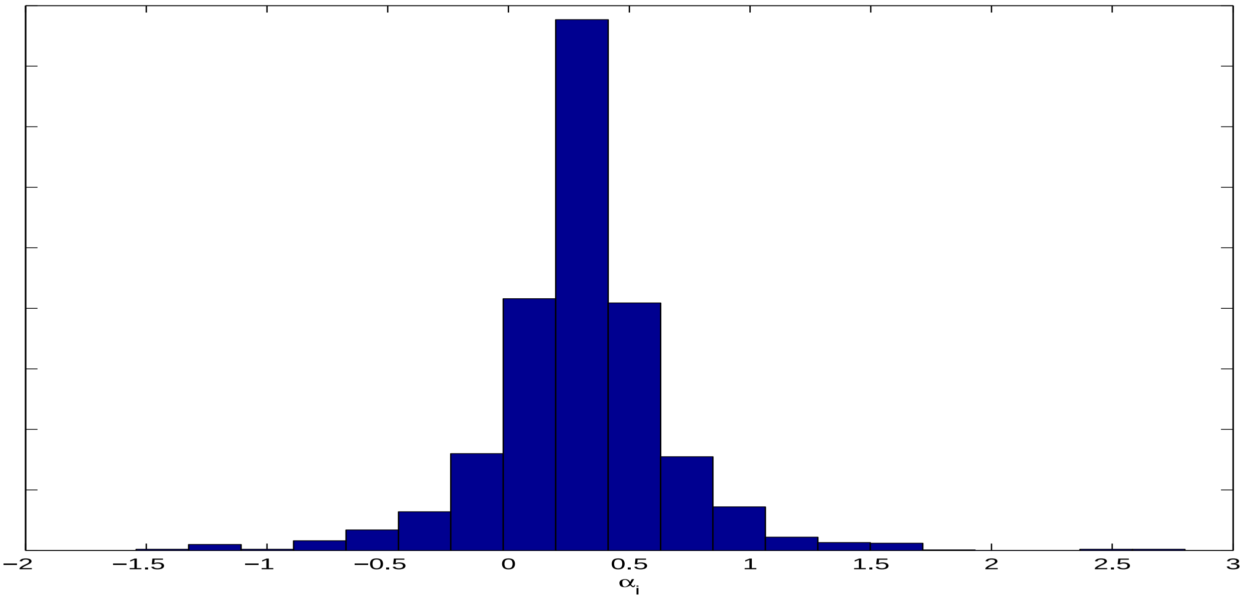


Figure 3: Histogram of Posterior Means of β_is for Model 2

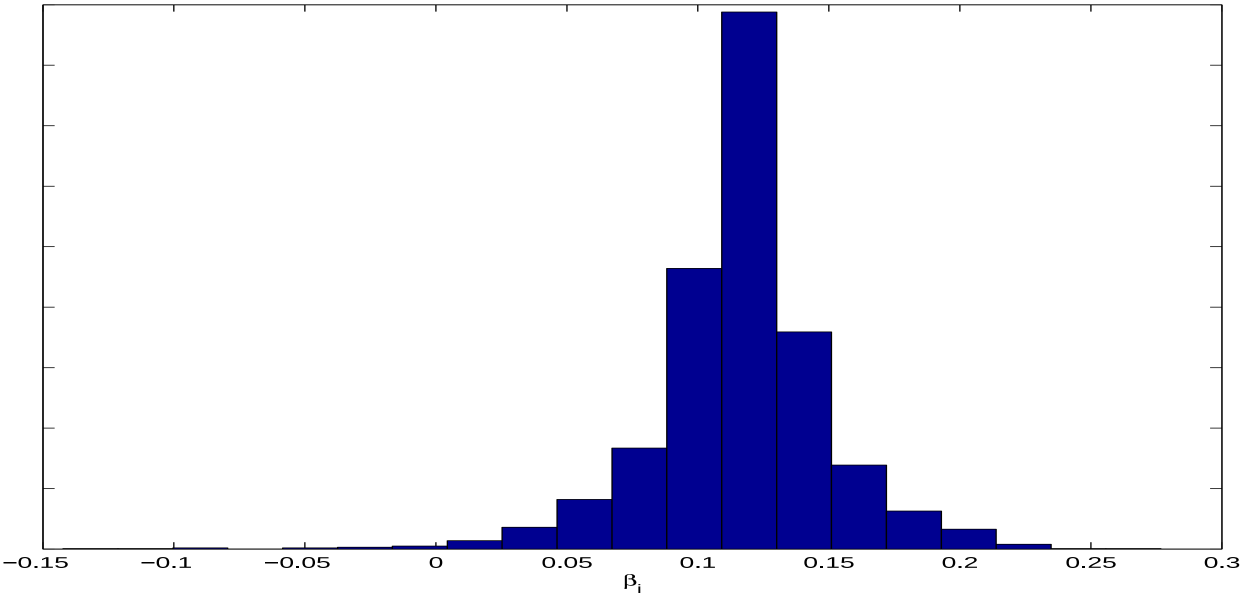
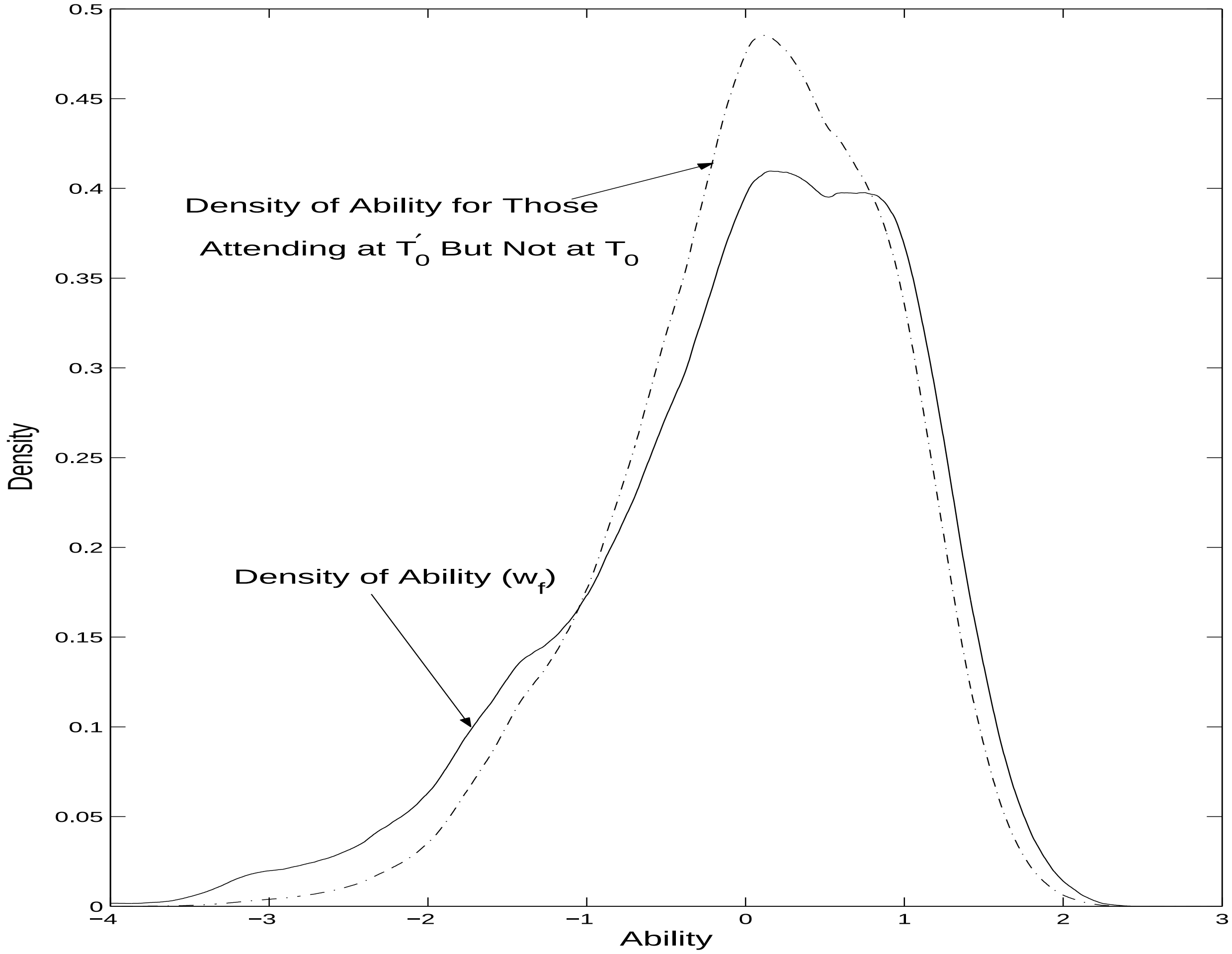


Figure 4: Nonparametric Estimates of the Population Distribution of Ability, and the Distribution of Ability for the"Compliers"



8 Appendix B: 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:

$$
\begin{aligned}
y_{it}|x_{it}, z_{it}, w_i, \theta_i, \gamma, \sigma_\epsilon^2 &\overset{ind}{\sim} N(x_{it}\theta_i + z_{it}\gamma, \sigma_\epsilon^2), \;\; i = 1, 2, \cdots, n, \;\; t = 1, 2, \cdots T_i. && (1)\\
\theta_i|\lambda, w_i &\overset{ind}{\sim} f(\theta_i|\lambda, w_i) && (2)\\
\gamma|\underline{\mu}_\gamma, \underline{V}_\gamma &\sim N(\underline{\mu}_\gamma, \underline{V}_\gamma) && (3)\\
\lambda|\underline{\lambda} &\sim g(\underline{\lambda}) && (4)\\
\sigma_\epsilon^{-2}|s_\epsilon^{-2}, \underline{\nu}_\epsilon &\sim G(\underline{s}_\varepsilon^{-2}, \underline{\nu}_\varepsilon), && (5)
\end{aligned}
$$

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, $G\left(\underline{s}_\varepsilon^{-2}, \underline{\nu}_\varepsilon\right)$ denotes the Gamma distribution with mean $\underline{s}_\varepsilon^{-2}$ and degrees of freedom $\underline{\nu}_\varepsilon$ (see Poirier, 1995, page 100),

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

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, $\theta = (\theta_1, .., \theta_N)'$,

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

and define y_i and Z_i conformably. Define

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

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.

Model 1: $\alpha_i = \alpha$ and $\beta_i = \beta$.

Let $\delta = (\alpha, \beta, \gamma')'$ then Model 1 is simply the Normal linear regression model:

$$y = U\delta + \varepsilon$$

where $U = [X\ Z]$. We use an independent Normal-Gamma prior for δ and σ_ε^{-2}. That is,

$$\delta \sim N(\underline{\delta}, \underline{V}) \tag{B.6}$$

and

$$\sigma_\varepsilon^{-2} \sim G(\underline{s}_\varepsilon^{-2}, \underline{\nu}_\varepsilon). \tag{B.7}$$

Note that (B.6) is (B.3) supplemented to include a Normal prior for α and β.

In this model, a Gibbs sampler can be set up in terms of the following conditional posteriors:

$$\delta | Data, \sigma_\varepsilon^{-2} \sim N\left(\overline{\delta}, \overline{V}\right). \tag{B.8}$$

and

$$\sigma_\varepsilon^{-2} | Data, \delta \sim G(\overline{s}_\varepsilon^{-2}, \overline{\nu}_\varepsilon), \tag{B.9}$$

where

$$\overline{V} = \left(\underline{V}^{-1} + \sigma_\varepsilon^{-2} U'U\right)^{-1}, \tag{B.10}$$

$$\overline{\delta} = \overline{V}\left(\underline{V}^{-1}\underline{\delta} + \sigma_\varepsilon^{-2} U'y\right), \tag{B.11}$$

$$\overline{\nu}_\varepsilon = NT + \underline{\nu}_\varepsilon \tag{B.12}$$

and

$$\overline{s}_\varepsilon^2 = \frac{(y - U\delta)'(y - U\delta) + \underline{\nu}_\varepsilon \underline{s}_\varepsilon^2}{\overline{\nu}_\varepsilon}. \tag{B.13}$$

The marginal likelihood for this model can calculated using the method outlined in Chib (1995). This requires the prior, likelihood and posterior to be evaluated at a point. The former two are straightforward. We evaluate the posterior using the fact that $p\left(\delta, \sigma_\varepsilon^{-2} | Data\right) = p\left(\delta | Data\right) p\left(\sigma_\varepsilon^{-2} | Data, \delta\right)$. The former of these can be evaluated inside the Gibbs loop (i.e. using (B.8) and Rao-Blackwellizing as described in Chib, 1995, page 1314-1315). The latter can be evaluated directly using (B.9).

Models 2 and 2W: $\theta_i | \theta_0, \Sigma, \tilde{w}_i \overset{ind}{\sim} N\left(\widetilde{w}_i \theta_0, \Sigma\right)$

In this hierarchical model, we define

$$\widetilde{w}_i = \begin{bmatrix} w_i & 0 \\ 0 & w_i \end{bmatrix}$$

and use a prior for θ_0 and Σ of the form:

$$\theta_0 \sim N(\underline{\theta}, \underline{V}_\theta) \tag{B.14}$$

and

$$\Sigma^{-1} \sim W\left(\left[\underline{\rho}\underline{\Sigma}\right]^{-1}, \underline{\rho}\right) \tag{B.15}$$

where $W(\cdot,\cdot)$ denotes the Wishart distribution defined as in Poirier (1995) page 136. The prior for the other parameters in given in (B.3) and (B.5). Note that Model 2 is the case where no explanatory variables appear in the second stage and $w_i = 1$. The obvious modification to $\widetilde{w}_i$ can be made if one wishes to have different explanatory variables for intercept and slope variation.

We employ a *blocking* or *grouping* step to reduce the autocorrelation in the parameter chains, and do this by drawing jointly from $\{\theta_i\}, \gamma|\Gamma_{-\gamma,\{\theta_i\}}$. Since the θ_i are conditionally independent, we draw from this joint distribution by first drawing γ from its conditional marginalized over $\{\theta_i\}$, and then drawing each θ_i independently from its complete conditional.

To this end, define $R_i \equiv X_i\Sigma X_i' + \sigma_\epsilon^2 I_{T_i}$, With this notation, the following conditionals are obtained:

$$\theta_i|Data, \Gamma_{-\theta_i} \overset{ind}{\sim} N(D_{\theta_i}d_{\theta_i}, D_{\theta_i}), \tag{B.16}$$

where $i = 1, .., N$,

$$D_{\theta_i} \equiv (X_i'X_i/\sigma_\epsilon^2 + \Sigma^{-1})^{-1}, \quad d_{\theta_i} \equiv X_i'(y_i - Z_i\gamma)/\sigma_\epsilon^2 + \Sigma^{-1}\widetilde{w}_i\theta_0.$$

$$\gamma|Data, \Gamma_{-\gamma,\{\theta_i\}} \sim N(D_\gamma d_\gamma, D_\gamma), \tag{B.17}$$

where

$$D_\gamma = (\sum_i Z_i'R_i^{-1}Z_i + \underline{V}_\gamma^{-1})^{-1}, \quad d_\gamma = \sum_i Z_i'R_i^{-1}(y_i - X_i\widetilde{w}_i\theta_0) + \underline{V}_\gamma^{-1}\underline{\mu}_\gamma.$$

Draw from the joint posterior conditional of $(\{\theta_i\}, \gamma)$ are obtained by drawing γ from (B.17), and then independently drawing each θ_i as in (B.16).

Next we have

$$\theta_0|Data, \Gamma_{-\theta_0} \sim N(D_{\theta_0}d_{\theta_0}, D_{\theta_0}), \tag{B.18}$$

where

$$D_{\theta_0} = \left[\widetilde{W}'\left(I_N \otimes \Sigma^{-1}\right)\widetilde{W} + \underline{V}_\theta^{-1}\right]^{-1}, \quad d_{\theta_0} = \widetilde{W}'\left(I_N \otimes \Sigma^{-1}\right)\theta + \underline{V}_\theta^{-1}\underline{\theta}.$$

where $\theta = (\theta_1', .., \theta_N')'$ and $\widetilde{W}$ is the $2N \times 2k_w$ matrix which stacks the $\widetilde{w}_i's$.

For the error precision we obtain

$$\sigma_\epsilon^{-2}|Data, \Gamma_{-\sigma_\epsilon^2} \sim G\left(\overline{s}_\varepsilon^{-2}, \overline{\nu}_\varepsilon\right) \tag{B.19}$$

where

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

and

$$\overline{s}^2_{\varepsilon} = \frac{\sum_i (y_i - X_i\theta_i - Z_i\gamma)'(y_i - X_i\theta_i - Z_i\gamma) + \underline{\nu}_{\varepsilon}\underline{s}^2_{\varepsilon}}{\overline{\nu}_{\varepsilon}}.$$

Finally we have,

$$\Sigma^{-1}|Data, \Gamma_{-\Sigma^{-1}} \sim W([\sum_i (\theta_i - \widetilde{w}_i\theta_0)(\theta_i - \widetilde{w}_i\theta_0)' + \underline{\rho}\underline{\Sigma}]^{-1}, N + \underline{\rho}). \tag{B.20}$$

The marginal likelihood can be calculated using the method of Chib (1995), page 1315. Note that the high dimensionality of the problem adds some slight complications. We can implement Chib's method by evaluating $p\left(\Gamma_{-\theta}\right), p\left(y|\Gamma_{-\theta}\right)$ and $p\left(\Gamma_{-\theta}|Data\right)$ at a particular value for $\Gamma_{-\theta}$ (e.g. the posterior mean based on a preliminary Gibbs run). The first and last of these can be evaluated at a point directly and through posterior simulation (see Chib, 1995, page 1315), respectively. The middle term is the likelihood function with θ integrated out (i.e. $p\left(y|\Gamma_{-\theta}\right) = \int p\left(y|\Gamma\right) p\left(\theta|\Gamma_{-\theta}\right) d\theta$). This can be calculated analytically (see, e.g., Lindley and Smith, 1972, pages 4-5) yielding $p\left(y|\Gamma_{-\theta}\right) = \prod_{i=1}^{N} p\left(y_i|\Gamma_{-\theta}\right)$ where

$$y_i|\Gamma_{-\theta} \;\sim\; N\left(X_i\widetilde{w}_i\theta_0 + Z_i\gamma, R_i\right).$$

Note also that $p\left(y|\Gamma_{-\theta}\right)$ is what we use in our calculation of the BIC. In practice, we use Chib's method for calculating the marginal likelihood for the case where $w_i = 1$. The Savage-Dickey density ratio is then used for calculating the Bayes factor comparing the model with $w_i = 1$ to the general case. In this way, the marginal likelihood for both models can be worked out.

Model 3 is simply a restricted version of Model 2 and the MCMC algorithm proceeds in the same manner.

Model 4: $\theta_i = \theta^0_g$ with probability π_g for $g = 1, .., G$.

This varies from a standard mixture model (e.g. Chib, 1995, pages 1318-1320) due to the fact that we are assuming γ and σ^2_{ε} are constant. Define the component label vector as $c_i \equiv (c_{1i}, .., c_{Gi})'$ with $c_{gi} = 1$ denoting that the i^{th} observation is drawn from the g^{th} 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

$$p(y|\Gamma) = \prod_{i=1}^{N} \left[\phi(y_i; X_i\theta^0_1 + Z_i\gamma, \sigma^2)\right]^{c_{1i}} \left[\phi(y_i; X_i\theta^0_2 + Z_i\gamma, \sigma^2)\right]^{c_{2i}} \cdots \left[\phi(y_i; X_i\theta^0_G + Z_i\gamma, \sigma^2)\right]^{c_{Gi}}, \tag{B.21}$$

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

$$p\left(c_i|\pi\right) = \prod_{g=1}^{G} \pi_g^{c_{gi}}, \tag{B.22}$$

where $\pi = (\pi_1, .., \pi_G)'$ 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:

$$\pi \sim Dir(a_1, a_2, , a_G), \tag{B.23}$$

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

$$\theta_g^0 \overset{ind}{\sim} N\left(\underline{\theta}_g, \underline{V}_g\right), \;\; \forall\, g = 1, 2, \cdots, G. \tag{B.24}$$

Posterior simulation is done using a Gibbs sampler with data augmentation involving the following posterior conditionals. First, for $g = 1, .., G$ we have

$$\theta_g^0 | Data, \Gamma_{-\theta_g} \overset{ind}{\sim} N\left(D_{\theta_g} d_{\theta_g}, D_{\theta_g}\right), \tag{B.25}$$

where

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

Next,

$$\sigma_\varepsilon^{-2} | Data, \Gamma_{-\sigma_\varepsilon^{-2}} \sim G(\overline{s}_\varepsilon^{-2}, \overline{\nu}_\varepsilon), \tag{B.26}$$

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

$$\gamma | Data, \Gamma_{-\gamma} \sim N(D_\gamma d_\gamma, D_\gamma), \tag{B.27}$$

where

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

Next we have, for $i = 1, .., N$,

$$c_i \overset{ind}{\sim} Mult\left(1, \left[\frac{\pi_1 \phi\left(y_i; X_i \theta_1^0 + Z_i \gamma, \sigma_\varepsilon^2 I_{T_i}\right)}{\sum_{g=1}^G \pi_g \phi\left(y_i; X_i \theta_g^0 + Z_i \gamma, \sigma_\varepsilon^2 I_{T_i}\right)}, .., \frac{\pi_G \phi\left(y_i; X_i \theta_G^0 + Z_i \gamma, \sigma_\varepsilon^2 I_{T_i}\right)}{\sum_{g=1}^G \pi_g \phi\left(y_i; X_i \theta_g^0 + Z_i \gamma, \sigma_\varepsilon^2 I_{T_i}\right)}\right]'\right), \tag{B.28}$$

where $Mult\left(\right)$ denotes the Multinomial distribution (see Poirier, 1995, page 118-119). Finally, we have

$$\pi | Data, \Gamma_{-\pi} \sim Dir(n_1 + a_1, n_2 + a_2, \cdots n_G + a_G), \tag{B.29}$$

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

Posterior simulation proceeds by sequentially simulating from (B.25), (B.26), (B.27), (B.28) and (B.29). The marginal likelihood can be calculated using the method of Chib (1995). Note that the high dimensionality of C adds some slight complications. However, we can implement Chib's method by evaluating $p\left(\Gamma_{-C}\right), p\left(y|\Gamma_{-C}\right)$ and $p\left(\Gamma_{-C}|Data\right)$ at a particular value for Γ_{-C} (e.g. the posterior mean based on a preliminary Gibbs run). The first and last of these can be evaluated at a point directly and through posterior simulation (see Chib, 1995, page 1319), respectively. The middle term is the likelihood function with C integrated out (i.e. $p\left(y|\Gamma_{-C}\right) = \int p\left(y|\Gamma\right) p\left(\theta|\Gamma_{-C}\right) dC$) which is given in Chib (1995, equation 16). When calculating BIC, we use $p\left(y|\Gamma_{-C}\right)$ as the likelihood function.

Model 4W: $\pi_{gi} = \pi_g(w_i)$.

This model extends Model 4 to allow explanatory variables to affect the heterogeneity. We let π_{gi} vary over individuals depending upon explanatory variables w_i and use an ordered probit specification (see, e.g., Albert and Chib, 1993) to try to explain the "mapping" between individuals and the mixture components. Model 4W is strictly a generalization of Model 4 - if the covariates w_i play no role in explaining the component assignment mechanism, then we are back into the constant probability of belonging to each component that is assumed in Model 3.

Following the ordered probit literature we work in terms of the c.d.f. and let

$$\eta_{gi} = \sum_{j=1}^{g} \pi_{ji}.$$

We then model

$$\eta_{gi} = \Phi\left(\tau_g - w_i'\zeta\right), \tag{B.30}$$

where $\Phi\left(a\right)$ is the c.d.f. of the standard Normal evaluated at point a, τ_j is a scalar parameter for $j = 1, ..., G-1$ with $\tau_1 < \tau_2 < .. < \tau_{m-1}$ and ζ is a k_w-vector of unknown coefficients. As discussed in, e.g., Albert and Chib (1993), we can impose, with no loss of generality, the identifying restriction $\tau_1 = 0$. To aid in interpretation, note that this model implies

$$\begin{aligned}
\pi_{1i} &= \Phi\left(-w_i'\zeta\right) \\
\pi_{2i} &= \Phi\left(\tau_2 - w_i'\zeta\right) - \Phi\left(-w_i'\zeta\right) \\
&\vdots \\
\pi_{G-1,i} &= \Phi\left(\tau_{G-1} - w_i'\zeta\right) - \Phi\left(\tau_{G-2} - w_i'\zeta\right) \\
\pi_{Gi} &= 1 - \Phi\left(\tau_{G-1} - w_i'\zeta\right)
\end{aligned}$$

An alternative way of interpreting this model is in terms of a latent variable. That is, let c_i^* be independently Normally distributed with mean $w_i'\zeta$ and standard deviation 1. This specification is equivalent to one where we observe $c_{ig} = 1$ if $\tau_{g-1} < c_i^* \leq \tau_g$ with $\tau_0 = -\infty$ and $\tau_G = \infty$. The latent data formulation provides the motivation for our Gibbs sampling with data augmentation algorithm described below. Such a formulation, of course, implies a hierarchical prior for C which is an extension of the one given in (B.22).

The prior for the parameters in this model as above (i.e. as specified in equations (B.3), (B.5) and (B.24)) with the addition of the hierarchical prior for the component indicators and:

$$\zeta \sim N(\underline{\zeta}, \underline{V}_\zeta). \tag{B.31}$$

Furthermore, for the elements of τ which are not restricted (remember that $\tau_0 = -\infty, \tau_1 = 0$ and $\tau_G = \infty$), we use a noninformative prior of the form:

$$p\left(\tau_g | \left\{\tau_h, h \neq g\right\}\right) = f_U\left(\tau_{g-1}, \tau_{g+1}\right), \tag{B.32}$$

where $f_U\left(\tau_{g-1}, \tau_{g+1}\right)$ is the Uniform density over the interval $\left(\tau_{g-1}, \tau_{g+1}\right)$.

The posterior simulator for doing inference in this model involves minor modifications of that developed previously. We know $p\left(\theta_g^0 | Data, \Gamma_{-\theta_g}\right) p\left(\gamma | Data, \Gamma_{-\gamma}\right)$ and $p\left(\sigma_\varepsilon^{-2} | Data, \Gamma_{-\sigma_\varepsilon^{-2}}\right)$ have the forms given in (B.25), (B.27) and (B.26), respectively, and $p\left(C | Data, \Gamma_{-C}\right)$ has form defined by (B.28) if we replace π_g by π_{gi} where the latter can be calculated based on (B.30). We also have the following complete conditional for ζ :

$$\zeta | Data, \Gamma_{-\zeta} = \overline{\zeta}, \overline{V}_\zeta, \tag{B.33}$$

where

$$\overline{V}_\zeta = \left(\underline{V}_\zeta^{-1} + W'W\right)^{-1}$$

and

$$\overline{\zeta} = \overline{V}_\zeta \left(\underline{V}_\zeta^{-1}\underline{\zeta} + W'C^*\right),$$

where $C^* = \left(c_1^*, .., c_N^*\right)'$.

The posterior conditionals for τ_g for $g = 2, .., G-1$ are given by:

$$p\left(\tau_g | Data, \Gamma_{-\tau_g}\right) = f_U\left(\overline{\tau}_{g-1}, \overline{\tau}_{g+1}\right), \tag{B.34}$$

where $h = 1, .., G$ and

$$\overline{\tau}_{g-1} = \max\left\{\max\left\{c_i^* : c_i = g\right\}, \tau_{g-1}\right\},$$

and

$$\overline{\tau}_{g+1} = \min\left\{\min\left\{c_i^* : c_i = g+1\right\}, \tau_{g+1}\right\}.$$

Finally, the c_i^*'s are conditionally independent over all i with

$$c_i^* | Data, \Gamma_{-c^*}, c_{gi} = 1 \sim TN_{(\tau_{g-1}, \tau_g]}\left(w_i \zeta, 1\right), \tag{B.35}$$

where $TN_{(a,b]}(\mu, \sigma^2)$ denotes the normal density with mean μ and variance σ^2 truncated to the interval $(a, b]$.Hence, a Gibbs sampler involving (B.25), (B.26), (B.27), (B.28), (B.33), (B.34) and (B.35) can be set up. For model comparison, we use the Savage-Dickey density ratio to compute Bayes factors involving restrictions on ζ.

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

Inference for this model follows similarly to that of Model 2 and Model 4 with slight modifications. For Model 5, the conditional likelihood function is:

$$p(y|\Gamma) = \prod_{i=1}^{n} \phi(y_i; X_i\theta_i + Z_i\gamma, \sigma^2), \tag{B.36}$$

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

$$\theta_i | \Gamma_{-\theta_i} = \left[\phi(\theta_i; \theta_1^0, \Sigma_1)\right]^{c_{1i}} \cdots \left[\phi(\theta_i; \theta_G^0, \Sigma_G)\right]^{c_{Gi}}, \tag{B.37}$$

and we choose priors for $\{\theta_g^0\}$, $c_i|\pi$ and π as in (B.24), (B.22) and (B.23), respectively. Finally, we put conjugate priors on the inverse covariance matrices Σ_g^{-1} as in (B.15) of Model 2:

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

We focus on forms of the complete posterior conditionals for this model which differ from those discussed in either Models 2 or 4:

$$\theta_i | \Gamma_{-\theta_i}, \text{Data} \overset{ind}{\sim} N(D_{\theta_i} d_{\theta_i}, D_{\theta_i}), \tag{B.38}$$

where

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

$$\theta_g^0 | \Gamma_{-\theta_g^0}, \text{Data} \overset{ind}{\sim} N(D_{\theta_g} d_{\theta_g}, D_{\theta_g}), \tag{B.39}$$

where

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

$$\Sigma_g^{-1} | \Gamma_{-\Sigma_g^{-1}}, \text{Data} \overset{ind}{\sim} W\left(\left[\sum_{i=1}^{n} c_{gi}(\theta_i - \theta_g^0)(\theta_i - \theta_g^0)' + \underline{\rho}_g \underline{\Sigma}_g \right]^{-1}, n_g + \underline{\rho}_g \right). \tag{B.40}$$

Throughout, we have defined $n_g \equiv \sum_{i=1}^{n} c_{gi}$. The posterior conditionals for c_i and π follow as in (B.28) and (B.29), where the second stage of the hierarchy takes the place of the conditional likelihood in (B.28). The posterior conditional for σ_ε^{-2} is equivalent to (B.19). The posterior conditional for γ is as given in (B.17) augmented to control for different components as in (B.27).

9 Appendix C: The Priors

Results are obtained using priors which select reasonable values for prior means, but then makes 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

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

Note that the resulting prior is completely noninformative for the error variance (and the value for $\underline{s}_\varepsilon^2$ is irrelevant) [32] 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

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

These hyperparameter values (which we also use for comparable parameters in other models), are sensible in light of previous empirical work. Specifically, our priors imply that on average, an added year of education increases hourly wages by 10 percent, and the intercept is centered so as to imply an individual with 10 years of schooling has an expected wage of around \$7, but prior variances are

[32] 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.

large so relatively little weight is attached to these prior guesses. All coefficients are assumed, *a priori*, to be independent of one another. This defines the hyperparameters labelled $\underline{\delta}, \underline{V}$ in Appendix B for Model 1. The fully noninformative variant sets $\underline{V}^{-1} = 0$.

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 centering 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 $\underline{V}_{\theta}^{-1}$. 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 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 1^{st} 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 center the prior over Model 2 - where the explanatory variables (except for the intercept) in w have no effect on the heterogeneity distribution, though prior variances imply relative noninformativeness.

To complete the prior for Model 4, we specify $\underline{\theta}_g$ and $\underline{V}_g$ as in (C.1), for all g. For the component probabilities, we make the noninformative choice of $a_1 = a_2 = \cdots = a_G = 1$. To complete the prior for Model 4W (where explanatory variables are added to explain the “mapping” of individuals to various components of the mixture model), we make the relatively noninformative choices of $\underline{\zeta} = 0_{k_w}$ and $\underline{V}_{\zeta} = I_{k_w}$. For the fully noninformative variants we set $\underline{V}_g^{-1} = 0$ and $\underline{V}_{\zeta} = 0$.

For Model 5, we use a prior which combines the ideas of Models 3 and 4. Thus, priors for θ_g^0 and Σ_g^{-1} for $g = 1, .., G$ are independent of one another, each identical to those used for θ and Σ^{-1} in Model 3.

The prior for π in Models 5 and 4 is identical.

For Models 4 and 5, we also impose the identifying restriction $\beta_1^0 < \beta_2^0 < \cdots < \beta_G^0$ through the prior. Note that this means that, for these models, the marginal likelihood calculated using the method of Chib (1995) will be an approximation since the truncation in prior and posterior is ignored.[33] However, since this truncation is only in one dimension of a high dimensional model and results obtained are consistent with BIC results, the approximation error is likely very small. Furthermore, if we do not impose identification, the Chib method will provide exact results. We have re-run our programs for $G = 2$ and 3 without imposing identification. Marginal likelihood results calculated using the Chib method are virtually the same regardless of whether the identifying restriction is imposed for these cases.

[33] The necessary integrating constants could be calculated using simulation methods. However, this would add greatly to an already substantive computational burden.

References

[1] Albert, J. and Chib, S. (1993). "Bayesian analysis of binary and polychotomous response data," *Journal of the American Statistical Association,* 88, 669-679.

[2] Allenby, G., Arora, N. and J. Ginter (1998). On the heterogeneity of demand," *Journal of Marketing Research*, forthcoming.

[3] Allenby, G.M. and Rossi, P. (1999). "Marketing models of consumer heterogeneity," *Journal of Econometrics,* 89, 57-78.

[4] Angrist, J. (1995). "The economic returns to schooling in the West Bank and Gaza Strip," *American Economic Review*, 85, 1065-1087.

[5] Angrist, J., Imbens, G. and Rubin, D. (1996). "Identification of causal effects using instrumental variables," *Journal of the American Statistical Association,* 91, 444-455.

[6] Angrist, J. and Krueger, A. (1991). "Does compulsory school attendance affect schooling and earnings?" *Quarterly Journal of Economics*, 106, 979-1014.

[7] Arias, O., Hallock, K. and Sosa-Escudero, W. (2001). "Individual heterogeneity in the returns to schooling: Instrumental variables quantile regression using twins data," *Empirical Economics,* 26, 7-40.

[8] Ashenfelter, O. and Krueger, A. (1994). "Estimates of the economic return to schooling from a new sample of twins," *American Economic Review*, 84, 1157-1173.

[9] Becker, G. and Chiswick, B. (1966). "Education and the distribution of earnings," *American Economic Review,* 56, 358-369.

[10] Ashenfelter, O. and J. D. Mooney (1968). "Graduate education, ability and earnings," *Review of Economics and Statistics* 50(1), 78-86.

[11] Belman, D. and Heywood, J. (1991). "Sheepskin effects in the returns to education: An examination of women and minorities," *Review of Economics and Statistics,* 73, 720-724.

[12] Blackburn, M. and Neumark, D. (1993). "Omitted-ability bias and the increase in the return to schooling," *Journal of Labor Economics* 11, 521-544.

[13] Blackburn, M. and Neumark, D. (1995). "Are OLS estimates of the return to schooling biased downward? Another look," *Review of Economics and Statistics,* 77, 217-230.

[14] Bound, J. and Jaeger, D. (1996). "On the validity of season of birth as an instrument in wage equations: A comment on Angrist and Krueger's 'Does compulsory school attendance affect schooling and earnings?', " *NBER Working Paper # 5835.*

[15] Bound, J., Jaeger, D. and Baker, R. (1995). "Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variables is weak," *Journal of the American Statistical Association,* 90, 443-450.

[16] Carlin, B and Polson, N. (1991). "Inference for nonconjugate Bayesian models using the Gibbs sampler," *Canadian Journal of Statistics,* 19, 399-405.

[17] Card, D. (2001). "Estimating the return to schooling: Progress on some persistent econometric problems," *Econometrica,* 69, 1127-1160.

[18] Cawley, J., Conneely, K., Heckman, J. and Vytlacil, E. (1997). "Cognitive ability, wages and meritocracy," in *Intelligence, Genes and Success: Scientists Respond to the Bell Curve.* B. Devlin, S.E. Feinberg, D.P. Resnick, and K. Roeder eds., New York: Springer, 179–192.

[19] Chib, S. (1995). "Marginal likelihood from the Gibbs output," *Journal of the American Statistical Association*, 90, 1313-1321.

[20] Chib, S. and B. Carlin (1999). "On MCMC sampling in hierarchical longitudinal data models," *Statistics and Computing,* 65, 361-394.

[21] Chiswick, B. (1974). *Income Inequality: Regional Analyses Within A Human Capital Framework.* New York: Columbia University Press.

[22] DiNardo, J. and Tobias, J. (2001). "Nonparametric density and regression estimation," *Journal of Economic Perspectives*, 15, 11-28.

[23] Escobar, M. and West, M. (1995). "Bayesian density estimation using mixtures," *Journal of the American Statistical Association*, 90, 577-588.

[24] Geweke, J. (1993). "Bayesian treatment of the independent Student t linear model," *Journal of Applied Econometrics,* 8, 19-40.

[25] Gilks, W., Richardson, S. and Spiegelhalter, D. (eds.). (1996). *Markov Chain Monte Carlo in Practice*, Boca Raton: Chapman and Hall.

[26] Grogger, J. and Eide, E. (1995). "Changes in college skills and the rise in the college wage premium," *Journal of Human Resources,* 30, 280-310.

[27] Harmon, C. and Walker, I. (1999). "The marginal and average returns to schooling in the UK," *European Economic Review*, 43, 879-887.

[28] Hansen. W.L., Weisbrod, B.A. and Scanlon, W. (1970). “Schooling and earnings of low achievers,” *American Economic Review* 60, 409-418.

[29] Hause, J. (1972). “Earnings profile: Ability and schooling,” *Journal of Political Economy* 80, S108-S138.

[30] Heckman, J. (1997). “Instrumental variables: A study of implicit behavioral assumptions in making program evaluations,” *Journal of Human Resources,* 32, 441-462.

[31] Heckman, J., Layne-Farrar, A. and 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.

[32] Heckman, J. and Polachek, S. (1974). “Empirical evidence on the functional form of the earnings-schooling relationship,” *Journal of the American Statistical Association,* 69, 350-354.

[33] 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* New York: Ca.

[34] Heckman, J. and Singer, B. (1984). “A method for minimizing the impact of distributional assumptions in econometric models for duration data,” *Econometrica,* 52, 271-320.

[35] Heckman, J., J.L. Tobias and E. Vytlacil (2001). “Simple Estimators for Treatment Parameters in a Latent Variable Framework.” Revised version of *NBER Working Paper* #7950.

[36] Heckman, J. and 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.

[37] Heckman, J. and 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.

[38] Heywood, J. (1994). “How widespread are sheepskin returns to education in the U.S.?” *Economics of Education Review*, 13, 227-234.

[39] Hobert, J. and 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.

[40] Hungerford, T. and Solon, G. (1987). “Sheepskin effects in returns to education,” *Review of Economics and Statistics*, 69, 175-177.

[41] Ichino, A. and 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.

[42] Imbens, G. and Angrist, J. (1994). “Identification and estimation of local average treatment effects,” *Econometrica,* 62, 467-475.

[43] Jaeger, D. and Page, M. (1996). “Degrees matter: New evidence on sheepskin effects in returns to education,” *Review of Economics and Statistics,* 78, 733-740.

[44] Kling, J. (2001). “Interpreting instrumental variables estimates of the returns to schooling,” *Journal of Business and Economic Statistics,* 19, 358-364.

[45] Lam, D. and Schoeni, R.F. (1993). “Effects of family background on earnings and returns to schooling: evidence from Brazil” *Journal of Political Economy*, 101(4), 710-740.

[46] Lindley, D. and Smith, A.F.M. (1972). ”Bayes estimates for the linear model,” *Journal of the Royal Statistical Society, series B*, 34, 1-41.

[47] McLachlan and Peel (2000). *Finite Mixture Models.* New York: Wiley.

[48] Mincer, J. (1958). “Investment in human capital and personal income distribution,” *Journal of Political Economy,* 66, 281-302.

[49] Mincer, J. (1974). *Schooling, Experience and Earnings.* New York: Columbia University Press.

[50] Murnane, R., Levy, F. and Willett, J. (1995). “The growing importance of cognitive skills in wage determination,” *Review of Economics and Statistics,* 77, 251-266.

[51] Poirier, D. (1995). *Intermediate Statistics and Econometrics.* Cambridge: MIT Press.