

# Ordered Discrete Choice Models with Stochastic Thresholds<sup>1</sup>

Karsten T. Hansen<sup>2</sup>, James J. Heckman<sup>3</sup> and Kathleen J. Mullen<sup>4</sup>

July 14, 2003

<sup>1</sup>This research is supported by grants from NICHD-40-4043-000-85-261 and NSF SES-0099195 and the American Bar Foundation.

<sup>2</sup>Address: Department of Economics, University of Chicago, 1126 E. 59th Street, IL 60637. E-mail: khansen@midway.uchicago.edu.

<sup>3</sup>Address: Department of Economics, The University of Chicago, 1126 East 59th Street, Chicago, IL 60637, and The American Bar Foundation. E-mail: jjh@uchicago.edu.

<sup>4</sup>Address: Department of Economics, The University of Chicago, 1126 East 59th Street, Chicago, IL 60637. E-mail: kjmullen@midway.uchicago.edu.

# 1 An Economic Model of Schooling Choice

This paper extends the ordered discrete choice model for schooling introduced in Cameron and Heckman (1998). Let  $S = s$  denote the individual's choice of schooling level, where  $s \in \{1, \dots, \bar{S}\}$ . Let  $c(s|x)$  denote the direct costs of attaining schooling level  $s$  conditional on individual background characteristics  $X = x$ . Assume that costs are increasing and weakly convex:  $c(s+1|x) - c(s|x) \geq c(s|x) - c(s-1|x)$ . Let  $R(s)$  denote the discounted lifetime return in utility terms to schooling level  $s$ . Assume that  $R$  is increasing and concave.  $R$  implicitly includes both postponed and forgone earnings. Individuals solve the following maximization problem:

$$\max_{s \in \{1, \dots, \bar{S}\}} \{R(s) - c(s|x)\}. \quad (1)$$

The assumptions on the return and cost functions ensure that net returns are concave in  $s$ ; ignoring ties, the optimal solution for the schooling level will be unique.

We introduce omitted variables into the model with two distinct ways. First we introduce a scalar random variable  $\epsilon$  representing an individual-specific shifter of the return relative to the cost that is observed and acted

upon by the individual but that is not observed by the econometrician. Note that  $\epsilon$  shifts the cost function by the same proportion across schooling levels. We assume that the cost function depends on observed and unobserved individual characteristics in a multiplicatively separable way,

$$c(s|x) = c(s|w_s)\varphi(z)\epsilon, \quad (2)$$

where  $x$  is partitioned into  $(w_s, z)$ ,  $\epsilon \geq 0$  and  $E(\epsilon) = 1$ . The  $z$  variables are common across all schooling levels. Define  $\nu_{s-1} = c(s|w_s) - c(s-1|w_{s-1})$ , the increase in direct costs in going from schooling level  $s-1$  to  $s$ . Assume that the individual knows and acts on  $\nu_s$  but conditional on the observables  $w_s$ , it is unobservable to the analyst. Convexity of  $c(s|x)$  implies  $\nu_s \geq \nu_{s-1}$ . Assume that the  $\nu_s$  terms are dependent in the following way:  $\nu_s = \psi_s \nu_{s-1}$  where  $\psi_s \geq 1$  and  $\psi_s \perp\!\!\!\perp \psi_{s'}$  for all  $s \neq s'$ .

At the optimal schooling level  $s$ , net returns must be at least as large as net returns at  $s-1$  and  $s+1$ :

$$R(s) - c(s|w_s)\varphi(z)\epsilon \geq R(s-1) - c(s-1|w_{s-1})\varphi(z)\epsilon$$

$$R(s) - c(s|w_s)\varphi(z)\epsilon \geq R(s+1) - c(s+1|w_{s+1})\varphi(z)\epsilon$$

Rearranging, we obtain the following inequalities:

$$\begin{aligned} \frac{R(s+1) - R(s)}{[c(s+1|w_{s+1}) - c(s|w_s)]\varphi(z)} &\leq \epsilon \leq \frac{R(s) - R(s-1)}{[c(s|w_s) - c(s-1|w_{s-1})]\varphi(z)} \\ \frac{R(s+1) - R(s)}{\nu_s \varphi(z)} &\leq \epsilon \leq \frac{R(s) - R(s-1)}{\nu_{s-1} \varphi(z)} \end{aligned}$$

Letting  $\varphi(z) = \exp(-z\gamma)$ ,  $U = -\log(\epsilon)$ ,  $\exp[-l(s)] = R(s+1) - R(s)$ ,

$v_s = \log(\nu_s)$ , we have:

$$l(s-1) + \nu_s \leq z\gamma + U \leq l(s) + \nu_{s+1}$$

Let  $C_s = C_{s-1} + \Delta_s$  and  $\Delta_s = l(s) - l(s-1) + \log(\psi_s)$ . Note that  $\nu_s - \nu_{s-1} =$

$\log(\psi_s \nu_{s-1}) - \log(\nu_{s-1}) = \log(\psi_s) \geq 0$  so that

$$\Delta_s = l(s) - l(s-1) + \log(\psi_s). \quad (3)$$

With this notation, the condition for choosing level  $s$  is

$$S = s \iff C_{s-1} \leq z\gamma + U < C_s, \quad (4)$$

and the assumptions above ensure that  $C_{s-1} < C_s$ . The interpretation of the

cut-off  $C_s$  follows by noting that we can write

$$\begin{aligned}
 C_s &= \log \left[ \frac{MR(s)}{MC(s)} \right] \\
 MR(s) &\equiv R(s) - R(s-1) \\
 MC(s) &\equiv c(s|w_s) - c(s-1|w_{s-1})
 \end{aligned}$$

Given that the individual has already completed schooling level  $s-1$ , the probability that he stops at schooling level  $s$  (as opposed to advancing on to a higher schooling level,  $S \geq s+1$ ) depends on the ratio of marginal revenue to marginal cost of advancement.

This model extends the classic ordered probit model in two ways: (a) thresholds depend on regressors and (b) thresholds are stochastic. We establish nonparametric identification of this model. See appendix A. The full model specification is completed by specifying the joint distribution of  $U$  and the sequence of MR/MC ratios  $C_s$ ,  $s \in S$ .<sup>1</sup> Appendix A established semi-parametric identification. In the next section, we consider a joint model of schooling and test scores.

---

<sup>1</sup>Mullen (2001) estimates a parametric version of this schooling model on NLSY data.

## 2 A Joint Model of Schooling and Test Scores

Let the index in the schooling model be  $V$ , i.e.,

$$V = z\gamma + U. \tag{5}$$

We will model the “cut-offs” as

$$C_s = C_{s-1} + \Delta_s(W_s, \widetilde{\Delta}_{ji}), \tag{6}$$

where the increment  $\Delta_s \geq 0$  is a function of observed ( $W_s$ ) covariates and unobservable ( $\widetilde{\Delta}_s$ ) components. Note that by construction  $C_s \geq C_{s-1}$ . For identification purposes we set  $C_0 \equiv -\infty$  and  $C_{\bar{s}} \equiv +\infty$ .

The index  $V$  can be thought of as a latent utility index where the  $Z$  variables (e.g. family background characteristics) unilaterally affect one’s probability of attending a given level of schooling. As was shown above, the cutoffs can be interpreted as net marginal benefit-cost ratios of transitioning to the next level of schooling, given that the individual has completed the prerequisite level. In the application of schooling choice, the ordering of the outcomes is clear: Individuals may not transition to the higher level (e.g.

college) without first completing all levels up to that point (e.g. elementary, secondary educations). The need for the ordering of the cutoffs is also clear: Having completed one level of schooling, an individual cannot return to the previous level of schooling.

Assuming separability of observables and unobservables we write

$$\log \Delta_s = W_s \delta_s + \tilde{\Delta}_s. \quad (7)$$

We partition the unobservable characteristics in the following way. Let  $f$  be a latent factor (e.g. an unobserved characteristic such as cognitive ability) such that conditional on  $f$  and on the observed regressors, all other unobservables are independently distributed across individuals and across transitions. Assume that

$$\tilde{\Delta}_s = \alpha_s f + \psi_s, \quad (8)$$

where  $\psi_s$  is standard normal. In this paper we restrict our analysis to a log normal parameterization of the model, where the increments  $\Delta_s$  are distributed log-normally conditional on  $f$ . However, we will estimate the distribution of  $f$  using a flexible mixtures of normal distribution so the uncondi-

tional distribution is practically non-parametric. The conditional log-normal specification is a natural choice in that it automatically restricts the innovations in the costs to be positive. In addition, in this paper we adopt a Bayesian framework for estimation and we use Markov Chain Monte Carlo (MCMC) techniques to sample from the joint and marginal posterior distributions of the model parameters conditional on the observed data; in particular we use an algorithm based on the Gibbs sampler, which samples iteratively from the posterior distributions of the parameters conditional on the data and on values for the other parameters. A log normal specification for the distribution of the cost innovations leads to computationally convenient conditional distributions of the model parameters.

Assume the unobservable component in the choice index can be decomposed as

$$U = \alpha_1 f + \varepsilon, \tag{9}$$

where  $\varepsilon$  is standard normal. The choice system can then be summarized as

follows:

$$V = Z\gamma + \alpha_1 f + \varepsilon$$

$$\varepsilon|f \sim N(0, 1)$$

$$C_1 \equiv 0$$

$$C_s = C_{s-1} + \Delta_s, \quad s = 2, \dots, \bar{S} - 1$$

$$\log \Delta_s = W_j \delta_j + \alpha_j f + \psi_s,$$

$$\psi_s|W, f \sim N(0, 1)$$

$$D = s \iff C_{s-1} \leq V < C_s$$

$$C_0 = -\infty, \quad C_{\bar{S}} = +\infty$$

We now turn to modeling the test scores. The raw AFQT score is defined as the sum of four sub-test components,

$$T = \sum_{j=1}^4 T_j,$$

where  $T_1$  is word knowledge,  $T_2$  is paragraph comprehension,  $T_3$  is arithmetic reasoning and  $T_4$  is math knowledge. A important feature of the ASVAB test components in the NLSY data is that a substantial number of test score

observations “hit the ceiling”, i.e., they are equal to the maximum score on a particular test component. This is documented in table 3 (see the data description below). To account for these ceiling effects we will work with a latent test score  $T_j^*$  so that

$$T_j = \begin{cases} T_j^* & \text{if } T_j^* < c_j, \\ c_j & \text{if } T_j^* \geq c_j, \end{cases} \quad (10)$$

where  $c_j$  is the maximum attainable score on test component  $j$ . Let the latent test score for an individual with schooling level  $s$  at the test date be

$$T^* = X_s \beta_s + \lambda_s f + \varepsilon_{T,s}, \quad (11)$$

where  $X_s$  is a set of observed covariates and  $f$  is the unobservable factor. We assume that –conditional on  $f$  –  $\varepsilon_{T,s}$  follows a normal distribution.

### 3 Semiparametric Identification of The Ordered Choice Model With Stochastic Thresholds

In the text, we write  $S = s$  if  $c_{s-1} \leq Z\gamma + U < c_s$  where  $c_s = c_{s-1} + \Delta_s$ ,  $\Delta_s \geq 0$ .

We normalize  $U = 0$ . It can be added or subtracted from the  $c_s$  without affecting the model. We write

$$\Delta_s = \varphi(w_s) + \nu_s$$

where  $E(\nu_s) = 1$  for all  $s$ .

We assume all variables have finite means and variances.

$$(A-1) \quad \nu_s \perp\!\!\!\perp \nu_{s'} \text{ all } s \neq s'$$

$$(A-2) \quad \nu_s \perp\!\!\!\perp (Z, w_1, \dots, w_{\bar{S}})$$

$$(A-3) \quad \varphi(w_1) = 1 \text{ (normalization)}$$

$$(A-4) \quad \nu_s, s = 1, \dots, \bar{S} \text{ are absolutely continuous}$$

$$(A-5) \quad 0 < Var(\nu_s) < \infty$$

$$(A-6) \quad Z \text{ has at least one continuous coordinate conditional on}$$

$$(W_1, \dots, W_{\bar{S}})$$

$$(A-7) \quad \text{Var}(\nu_1) = 1$$

$$(A-8) \quad W_s \text{ conditional on } W_{s-1}, \dots, W_1, Z \text{ is variation free}$$

**Theorem:** Under assumptions (A-1) - (A-8)  $\nu_1 \varphi(W_j), j = 1, \dots, \bar{S}$  and the densities of  $\nu_1, \dots, \nu_{\bar{S}}$  are identified.

Proof

First assume that  $\varphi(w_s) = k_s$  for all  $s$ . Then we may write

$$\Pr(S = 1 \mid Z) = \Pr(Z\gamma - \nu_1 \leq k_1)$$

so from standard results (see *e.g.* Cameron and Heckman, 1998) we can identify the density  $f(\nu_1)$  up to scale (*i.e.* we can identify  $k_1|\sigma_1, \gamma/\sigma_{\nu_1}$  and  $f\nu_1/\sigma_1$ ).

Now observe that

$$\Pr(S \leq 2 \mid Z) = \Pr(Z\gamma < k_2 + \nu_1 + \nu_2).$$

Let  $\sigma_{12} = \sqrt{\sigma_{\nu_1}^2 + \sigma_{\nu_2}^2}$ . Therefore by the same reasoning as previously used,

we can identify  $\gamma/\sigma_{12}$ ,  $c_2/\sigma_{12}$  and  $\frac{k_1 + k_2}{\sigma_{12}}$  and the density

$$f((\nu_1 + \nu_2)/\sigma_{12}). \tag{A-1}$$

Thus we can combine results and estimate

$$\frac{\sigma_{\nu_1}^2 + \sigma_{\nu_2}^2}{\sigma_{\nu_1}^2}.$$

Normalizing  $\sigma_{\nu_1}^2 = 1$ , we can identify  $\sigma_{\nu_2}^2$  and  $k_2$ . From the density (a-1) we can construct the characteristic function  $\psi$

$$\psi\left(\frac{\nu_1 + \nu_2}{\sigma_{12}}\right) = \psi\left(\frac{\nu_1}{\sigma_{12}}\right) \psi\left(\frac{\nu_2}{\sigma_{12}}\right).$$

From the first step, we identify

$$\psi\left(\frac{\nu_1}{\sigma_1}\right)$$

where  $\sigma_1 = [Var(\nu_1)]^{1/2} = 1$ . We know  $\sigma_{12}$  and hence

$$\psi\left(\frac{\nu_2}{\sigma_{12}}\right) = \frac{\psi\left(\frac{\nu_1 + \nu_2}{\sigma_{12}}\right)}{\psi\left(\frac{\nu_1}{\sigma_{12}}\right)}$$

from the inversion theorem we can identify the density of  $\nu_2$  (subject to the normalization of  $\nu_1 = 1$ ). We identify the constant  $k_2$  up to scale ( $\sigma_1$ ).

Proceeding sequentially, we can identify  $\sigma_{1,2,3,4,\dots,j} = \sqrt{\sigma_{\nu_1}^2 + \dots + \sigma_{\nu_{\bar{S}}}^2}$ ,  $j = 1, \dots, \bar{S} - 1$  (hence all of the  $\sigma_{\nu_j}^2$  given the normalization  $\sigma_{\nu_1}^2 = 1$ ), and the densities  $f(\nu_j)$ ,  $j = 1, \dots, \bar{S} - 1$  and the constants  $k_1, \dots, k_{\bar{S}-1}$ .

Observe that we obtain identification of the normalized  $\gamma$ , the  $\sigma_{\nu_s}^2$ , and the densities  $f(\nu_1), \dots, f(\nu_{\bar{S}-1})$ , without any exclusion restrictions. Suppose now that we allow  $\varphi(w_s)$  to be nontrivial functions of  $w_s$ ,  $s = 1, \dots, \bar{S} - 1$ . We may write

$$\Pr(S = 1 \mid Z) = \Pr(Z\gamma < C(W_1)\nu_1).$$

Conditional on  $W_1$ , as a consequence of (A-7), we can identify the intercept  $\varphi(w_1)$  for all values of the  $W_1$  in the support.

Applying (A-7) sequentially, we may write

$$\Pr(S = 2 \mid Z) = \Pr(\varphi_1(w_1) + \nu_1 \leq Z\gamma\varphi_2(W_2) + \nu_1 + v_2).$$

Note that  $\gamma, \varphi_1(W_1)$  are determined by the preceding arguments. Proceeding in this fashion, we may identify  $\varphi(W_2), \dots, \varphi(W_{\overline{S}-1})$ . Note that we can substitute  $U$  for  $\nu_1$  (i.e. set  $\nu_1 \equiv 0$  and  $U \neq 0$ ).

An alternative model replaces (A-1) with the factor assumption

$$\nu_s = \alpha_s f + \tau_s \tag{A-1}$$

where  $\tau_s$  are mean zero, mutually independent and identically distributed random variables with  $f \perp\!\!\!\perp (Z, W_1, \dots, W_{\bar{S}})$ .

Proof: Initially fix  $\varphi(W_s) = k_s, s = 1, \dots, \bar{S}^{-1}$ , a sequence of constants. We normalize  $\text{Var}(\tau_s) = 1$ , (or some other known numbers) for all  $s, \sigma_f^2 = 1$  and  $\alpha_1 = 1$ . Proceeding sequentially, we obtain identification

$$g\left(\frac{\nu_1}{\sigma_1}\right) = g\left(\frac{\alpha_1 f + \tau_1}{\sigma_1}\right)$$

where  $\sigma_1^2 = \alpha_1^2 \sigma_f^2 + \sigma_\tau^2 = 2$ . Using deconvolution, we can identify the density of  $f(g) f / \sqrt{2}$  if we assume a functional form for  $g(\tau_1 / \sqrt{2})$ . From

$$\Pr(S \leq 2 | Z\gamma \leq (k_2 + v_1 + v_2)) =$$

$$\Pr(S \leq 2 | Z\gamma \leq k_2 + (\alpha_1 + \alpha_2)f + \tau_1 + \tau_2)$$

we can identify

$$\gamma/\sigma_{\nu_1+\nu_2}$$

where

$$\sigma_{\nu_1+\nu_2}^2 = (\alpha_1 + \alpha_2)^2 \sigma_f^2 + 2\sigma_\tau^2.$$

Recall that we can identify  $\gamma/\sigma_{\nu_1}$  from the preceding argument. Hence we can identify

$$\frac{\sigma_{\nu_1+\nu_2}^2}{\sigma_{\nu_1}^2} = \frac{(1 + 2\alpha_2 + \alpha_2^2)\sigma_f^2 + 2\sigma_\tau^2}{\sigma_f^2 + \sigma_\tau^2} \quad (\text{A-2})$$

Given  $\sigma_\tau^2 = \sigma_f^2 = 1$ , we can identify  $\alpha_2$  up to sign,

$$-1 \pm \left[ 2 \left( \frac{\sigma_{\nu_1+\nu_2}^2 - \sigma_{\nu_1}^2}{\sigma_{\nu_1}^2} \right) \right]^{1/2} = \alpha_2.$$

Proceeding sequentially we can identify  $\alpha_2, \dots, \alpha_{\bar{S}-1}, g(f)$  and  $\frac{k_1}{\sqrt{2}}, \dots, \frac{k_{\bar{S}-1}}{\sqrt{2}}$ .

We can normalize  $\alpha_1, \sigma_f^2$  and  $\sigma_\tau^2$  in alternative ways. Proceeding as in the pre-

vious theorem, we can identify  $\frac{k_1}{\sqrt{2}}, \dots, \frac{k_{\bar{S}-1}}{\sqrt{2}}$  for each value of  $W_1, \dots, W_{\bar{S}-1}$   
and hence we can identify  $\frac{\varphi_1(W_1)}{\sqrt{2}}, \dots, \frac{\varphi_{\bar{S}-1}(W_{\bar{S}-1})}{\sqrt{2}}$ .

## 4 One Factor Model

### Cross Sectional Version

$$Y_1 = \mu_1 + \lambda_1 f + \varepsilon_1 \quad (12)$$

...

$$Y_j = \mu_j + \lambda_j f + \varepsilon_j$$

...

$$Y_N = \mu_N + \lambda_N f + \varepsilon_N$$

$N$  = Number of Outcomes

## Index of Choices

$$I = Z\eta + V \quad (13)$$

$$V = \gamma f + \varepsilon_I$$

$$\sigma_V^2 = \gamma^2 \sigma^2 f + \sigma_I^2$$

where  $E(\varepsilon_I^2) = \sigma_I^2$ .

$D_i = 1$  ( $Y_i$  observed) if

$$c_{i-1} < I \leq c_i$$

for  $i = 1, -\infty < I \leq c_1$

for  $i = N, c_{N-1} < I < \infty$

1. Assume initially joint normality

2.  $f \perp\!\!\!\perp (\varepsilon_1, \varepsilon_2, \varepsilon_3, \dots, \varepsilon_N, \varepsilon_I)$

$$E(Y_1 | I \leq c_1) = \mu_1 + \frac{(\lambda_1 \gamma) \sigma_f^2}{(\gamma^2 \sigma_f^2 + \sigma_I^2)^{1/2}} K(1)$$

where

$$K(1) = \frac{\frac{1}{\sqrt{2\pi}} \exp -\frac{1}{2} \left( \frac{c_1 - Z\eta}{\sigma_V} \right)^2}{\Phi \left( \frac{c_1 - Z\eta}{\sigma_V} \right)}.$$

The expression for  $K(N)$  is defined analogously.

For  $(N > j > 1)$ , we may write

$$E(Y_j | c_{j-1} < I \leq c_j) = \mu_j + \frac{(\lambda_j \gamma) \sigma_f^2}{(\gamma^2 \sigma_f^2 + \sigma_{\varepsilon_I}^2)^{1/2}} K(j)$$

where

$$K(j) = \left[ \frac{c_j - Z\eta}{\sigma_V} \int_{\frac{c_{j-1} - Z\eta}{\sigma_V}}^{\frac{c_j - Z\eta}{\sigma_V}} t \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt \right] / \left[ \Phi \left( \frac{c_j - Z\eta}{\sigma_V} \right) - \Phi \left( \frac{c_{j-1} - Z\eta}{\sigma_V} \right) \right]$$

$$K(j) = \frac{\frac{1}{\sqrt{2\pi}} \left[ \exp -\frac{1}{2} \left( \frac{c_{j-1} - Z\eta}{\sigma_V} \right)^2 - \exp -\frac{1}{2} \left( \frac{c_j - Z\eta}{\sigma_V} \right)^2 \right]}{\Phi \left( \frac{c_j - Z\eta}{\sigma_V} \right) - \Phi \left( \frac{c_{j-1} - Z\eta}{\sigma_V} \right)}$$

Under normality, we can identify

$$\frac{\eta}{\sigma_V}, \frac{c_1}{\sigma_V}, \frac{c_2}{\sigma_V}, \dots, \frac{c_{N-1}}{\sigma_V}$$

(Classical result; see Cameron and Heckman, 1998)

∴ We can identify

$$\frac{\lambda_1 \gamma \sigma_f^2}{\sigma_V}, \dots, \frac{\lambda_i \gamma \sigma_f^2}{\sigma_V}, \dots, \frac{\lambda_N \gamma \sigma_f^2}{\sigma_V} \tag{14}$$

Normalize  $\alpha_1 = 1$

∴ we can identify  $\alpha_i$ ,  $i = 2, \dots, N$ , from ratios of (3).

If we set  $\sigma_V^2 = 1$ , we can identify  $\gamma \sigma_f^2$ .

Can we identify  $\sigma_i^2$ ,  $i = 1, \dots, N$ ? Yes, under normality, and yes more

generally as we shall see.

Assume  $Y$  is  $N(\mu, 1)$ . Then

$$mgf(t) = E(e^{tY} | M_2 < Y < M_1) = \frac{\int_{M_2}^{M_1} e^{tY} \frac{1}{\sqrt{2\pi}} \exp - \frac{1}{2}(Y - \mu)^2 dY}{\int_{M_2}^{M_1} \frac{1}{\sqrt{2\pi}} \exp - \frac{1}{2}(Y - \mu)^2 dY}$$

Complete the square to get

$$\begin{aligned} &= \exp\left(\mu t + \frac{t^2}{2}\right) \frac{\int_{M_2}^{M_1} \frac{1}{\sqrt{2\pi}} \exp - \frac{1}{2}(Y - (\mu + t))^2 dY}{\int_{M_2}^{M_1} \frac{1}{\sqrt{2\pi}} \exp - \frac{1}{2}(Y - \mu)^2 dY} \\ &= \exp\left(\mu t + \frac{t^2}{2}\right) \frac{\Phi(M_1 - \mu - t) - \Phi(M_2 - \mu - t)}{\Phi(M_1 - \mu) - \Phi(M_2 - \mu)} \end{aligned}$$

where  $\Phi$  is the unit normal *cdf*.

$$\therefore \text{mean is } \left. \frac{\partial mgf}{\partial t} \right|_{t=0} = \mu + \frac{\frac{1}{\sqrt{2\pi}} \left[ \exp - \frac{(M_2 - \mu)^2}{2} - \exp - \frac{(M_1 - \mu)^2}{2} \right]}{\Phi(M_1 - \mu) - \Phi(M_2 - \mu)} = J + \mu$$

Variance is then

$$\left. \frac{\partial^2 mgf}{\partial t^2} \right|_{t=0} - \text{mean}^2$$

$$\frac{\partial mgf}{\partial t} = (\mu + t) \exp\left(\mu t + \frac{t^2}{2}\right) \frac{\Phi(M_1 - \mu - t) - \Phi(M_2 - \mu - t)}{\Phi(M_1 - \mu) - \Phi(M_2 - \mu)}$$

$$+ \exp\left(\mu t + \frac{t^2}{2}\right) \frac{\left(\frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}(M_2 - \mu - t)^2\right] - \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}(M_1 - \mu - t)^2\right]\right)}{\Phi(M_1 - \mu) - \Phi(M_2 - \mu)}$$

$$= \exp\left(\mu t + \frac{t^2}{2}\right)$$

$$\left\{ (\mu + t) \frac{[\Phi(M_1 - \mu - t) - \Phi(M_2 - \mu - t)]}{\Phi(M_1 - \mu) - \Phi(M_2 - \mu)} + \frac{\frac{1}{\sqrt{2\pi}} \exp - \frac{1}{2}(M_2 - \mu - t)^2 - \frac{1}{\sqrt{2\pi}} \exp - \frac{1}{2}(M_1 - \mu - t)^2}{\Phi(M_1 - \mu) - \Phi(M_2 - \mu)} \right\}$$

$$\left. \frac{\partial^2 mgf}{\partial t^2} \right|_{t=0} = \mu \{\mu + J\} + \{1 + \mu J\}$$

$$+ \frac{\frac{(M_2 - \mu)}{\sqrt{2\pi}} \exp - \frac{1}{2}(M_2 - \mu)^2 - \frac{(M_1 - \mu)}{\sqrt{2\pi}} \exp - \frac{1}{2}(M_1 - \mu)^2}{\Phi(M_1 - \mu) - \Phi(M_2 - \mu)}$$

$$\therefore \text{Var}(Y | M_2 < Y < M_1) = \mu^2 + \mu J + 1 + \mu J$$

$$+ \frac{\frac{(M_2 - \mu)}{\sqrt{2\pi}} \exp - \frac{1}{2}(M_2 - \mu)^2 - \frac{(M_1 - \mu)}{\sqrt{2\pi}} \exp - \frac{1}{2}(M_1 - \mu)^2}{\Phi(M_1 - \mu) - \Phi(M_2 - \mu)}$$

$$- \mu^2 - 2\mu J - J^2$$

$$= 1 + \frac{\frac{(M_2 - \mu)}{\sqrt{2\pi}} \frac{1}{2}(M_2 - \mu)^2 - \frac{(M_1 - \mu)}{\sqrt{2\pi}} \exp - \frac{1}{2}(M_1 - \mu)^2}{\Phi(M_1 - \mu) - \Phi(M_2 - \mu)} - J^2$$

Define  $V(j)$  as

$$V(j) = 1 + \frac{\frac{(c_j - Z\eta)}{\sqrt{2\pi}} \exp -\frac{1}{2}(c_j - Z\eta)^2 - (c_{j-1} - Z\eta) \exp -\frac{1}{2} \frac{(c_{j-1} - Z\eta)^2}{2}}{\Phi(c_j - Z\eta) - \Phi(c_{j-1} - Z\eta)} - J^2(j)$$

||

Q

$$J(j) = \frac{\frac{1}{\sqrt{2\pi}} \exp -\frac{1}{2}(c_j - Z\eta)^2 - \frac{1}{\sqrt{2\pi}} \exp -\frac{1}{2}(c_{j-1} - Z\eta)^2}{\Phi(c_j - Z\eta) - \Phi(c_{j-1} - Z\eta)}.$$

Then for each choice  $k = 2, \dots, N - 1$ , we get the variance of  $Y_j$  given  $D_j = 1$

as

$$Var(Y_j | D_j = 1) = (\gamma\alpha_j\sigma_f^2)^2(Q(j) - J(j))^2 + \alpha_j^2\sigma_f^2 + \sigma_j^2$$

using  $\sigma_V^2 = 1$ . We know  $(\gamma\alpha_j\sigma_f^2)$ , can estimate  $\alpha_j^2$ ,  $\sigma_f^2 + \sigma_j^2$  from errors of  $Y_j$

given  $D_j = 1$ .

$\therefore$  If  $\gamma = 1$ , we can identify the error variance  $\sigma_j^2$ . Otherwise, it is not

identified. Given exclusion restrictions, we have  $\mu_1, \dots, \mu_T$ , given  $\alpha_1 = 1$ , we

get  $\alpha_2, \dots, \alpha_N$ . Given  $\sigma_V^2 = 1$ , we get  $\gamma\sigma_f^2$ .

We know that

$$\sigma_V^2 = \gamma^2 \sigma_f^2 + \sigma_I^2$$

$\therefore$  if  $\sigma_V^2 = 1$  we know  $\sigma_I^2 = 1 - \gamma^2 \sigma_f^2$ . If we set  $\gamma = 1$ , we can identify  $\sigma_f^2$  and hence  $\sigma_j^2$ ,  $j = 1, \dots, T$ .

$\therefore$  Without the measurement equation, we identify the full model in the cross section with the normalizations  $\alpha_1 = 1$ ,  $\gamma = 1$ , and  $\sigma_V^2 = 1$  which imply that

$$\sigma_I^2 = 1 - \sigma_f^2$$

$\therefore$  there is a bound on  $\sigma_f^2$ . Technically, nothing is wrong with a negative variance for  $\sigma_I^2$  in forming the likelihood but it is strange.

## Mixtures of Normals Version

Suppose that we have an  $R$ -component normal that is mixing on  $f$

$$f \sim \sum_{i=1}^R P_i N(\mu_i, \sigma_i^2)$$

$$P_i \geq 0 \quad \sum_{i=1}^R P_i = 1.$$

We continue to assume that all of the  $\varepsilon_i$  (single or double subscripted) are normal.

Normalize so that

$$\sum_{i=1}^R P_i \mu_i = 0.$$

The case  $\sigma_i^2 = \sigma^2$  is much easier to analyze but we consider the general case here. Write  $f_i = \mu_i + \varepsilon_i$ .

The key insight for the mixture of normals case is that within each sub-

distribution,  $i, i = 1, \dots, R$ , we get standard normal results.

Thus for the subpopulation  $i$ , letting  $\sigma_{V_i}^2 = \gamma^2 \sigma_{f_i}^2 + \sigma_I^2$ , we obtain

$$E(f_i | c_{j-1} < I_i \leq c_j) = \mu_i + \left( \frac{\gamma \sigma_{f_i}^2}{\sigma_{V_i}} \right) E \left( \frac{\varepsilon_i}{\sigma_{V_i}} \mid \frac{c_{j-1}}{\sigma_{V_i}} < \frac{I_i}{\sigma_{V_i}} \leq \frac{c_j}{\sigma_{V_i}} \right) = \mu_i + \frac{\gamma \sigma_{f_i}^2}{\sigma_{V_i}} K_i(j)$$

where

$$K_i(j) = \frac{\left[ \frac{1}{\sqrt{2\pi}} \exp -\frac{1}{2} \left( \frac{c_j - Z\eta - \mu_i}{\sigma_{V_i}} \right)^2 - \frac{1}{\sqrt{2\pi}} \exp -\frac{1}{2} \left( \frac{c_{j-1} - Z\eta - \mu_i}{\sigma_{V_i}} \right)^2 \right]}{\left[ \Phi \left( \frac{c_j - Z\eta - \mu_i}{\sigma_{V_i}} \right) - \Phi \left( \frac{c_{j-1} - Z\eta - \mu_i}{\sigma_{V_i}} \right) \right]}$$

so

$$E(f) = \sum_{i=1}^R E(f_i | c_{j-1} < I_i \leq c_j) P_i$$

so we may write

$$E(Y_{jt} | c_{j-1} < I_i \leq c_j) = \mu_{jt} + (\alpha_{jt} \gamma) \sum \left( \frac{\sigma_{f_i}^2}{\sigma_{V_i}} \right) K_i(j) P_i$$

this is the mixtures of normals generalization of ( ). To get second moment

results note that we may write

$$Y_{jt} = \mu_{jt} + \alpha_{jt}f_i + \varepsilon_{jt} = \mu_{jt} + \alpha_{jt}\mu_i + (\alpha_{jt}\varepsilon_i + \varepsilon_{jt})$$

$$\begin{aligned} E(Y_{jt} Y_{j't'} \mid c_{j-1} < I_i \leq c_j) &= \mu_{jt}\mu_{j't'} + (\alpha_{jt}\mu_{j't'} + \alpha_{j't'}\mu_{jt})E(f_i \mid \\ c_{j-1} < I_i \leq c_j) &+ \alpha_{jt}\alpha_{j't'}E(f_i^2 \mid c_{j-1} < I_i \leq c_j) + E(\varepsilon_{jt}\varepsilon_{j't'} \mid c_{j-1} < I_i \leq \\ c_j). \end{aligned}$$

Observe that the final term is zero unless  $t = t'$ . We previously derived the expression for

$$E(f_i \mid c_{j-1} < I_i \leq c_j) = \mu_i + E(\varepsilon_i \mid c_{j-1} < I_i \leq c_j).$$

We need only derive the expression for

$$E(f_i^2 \mid c_{j-1} < I_i \leq c_j).$$

Recall that  $f_i = \mu_i + \varepsilon_i$ . Therefore we obtain

$$E(f_i^2 | c_{j-1} < I_i \leq c_j) = \mu_i^2 + 2\mu_i E(\varepsilon_i | c_{j-1} < I_i \leq c_j) + E(\varepsilon_i^2 | c_{j-1} < I_i \leq c_j).$$

From the results on the *mgf*, we obtain

$$E(\varepsilon_i^2 | c_{j-1} < I_i \leq c_j) = \sigma_{f_i}^2 + \left( \frac{\gamma \sigma_f^2}{\sigma_{V_i}} \right)^2 E \left[ \left( \frac{I_i}{\sigma_{V_i}} \right) \mid \frac{c_{j-1}}{\sigma_{V_i}} < \frac{I_i}{\sigma_{V_i}} \leq \frac{c_j}{\sigma_{V_i}} \right]$$

where

$$E \left[ \left( \frac{I_i}{\sigma_{V_i}} \right)^2 \mid \frac{c_{j-1}}{\sigma_{V_i}} < \frac{I_i}{\sigma_{V_i}} \leq \frac{c_j}{\sigma_{V_i}} \right] = \frac{1 + \frac{1}{\sqrt{2\pi}} t_{ji} \exp -\frac{1}{2} t_{ji}^2 - \frac{1}{\sqrt{2\pi}} t_{j-1,i} \exp -\frac{1}{2} (t_{j-1,i})^2}{\Phi(t_{ji}) - \Phi(t_{j-1,i})}$$

$$\text{where } t_{ji} = \frac{c_j - Z\eta - \gamma\mu_i}{\sigma_{V_i}}.$$

Putting all of this together by summing over the components of  $i$ , we

obtain

$$\begin{aligned} E(Y_{jt} Y_{j't'} | c_{j-1} < I \leq c_j) &= \mu_{jt} \mu_{j't'} + (\alpha_{jt} \mu_{j't'} + \alpha_{j't'} \mu_{jt}) \sum_{i=1}^R P_i E(f_i | c_{j-1} < \\ I_i \leq c_i) \\ &+ \alpha_{jt} \alpha_{j't'} \sum_{i=1}^R P_i \left\{ \mu_i^2 + 2\mu_i E(\varepsilon_i | c_{j-1} < I_i \leq c_j) + \sigma_{f_i}^2 + \frac{(\gamma \sigma_f^2)^2}{\sigma_{V_i}^2} E \left[ \left( \frac{I_i}{\sigma_{V_i}} \right)^2 \mid \frac{c_{j-1}}{\sigma_{V_i}} < \frac{I_i}{\sigma_{V_i}} \leq \frac{c_j}{\sigma_{V_i}} \right] \right\} \\ &+ E(\varepsilon_t \varepsilon_{t'}). \end{aligned}$$

Observe that the final term is zero if  $t \neq t'$ .

## 5 Nonparametric Identification

Cameron and Heckman (1998) present conditions for the nonparametric identifiability of the ordered probit model for discrete choice (2).

Here we extend the model by introducing spell-specific regressors  $c_i(x_i)$ ,  $i = 1, \dots, N$  where we assume that  $X_1, \dots, X_N$  are variation free conditional on  $Z$ . Their proof shows that if  $V$  is absolutely continuous with density  $g(v) > 0$  almost everywhere,  $V \perp\!\!\!\perp Z$  and  $Z$  is a  $J$  vector excluding the intercept with at least one coordinate  $Z_j$  continuously variable over  $R^1$  and the  $J - 1$  vector  $\tilde{Z}_m$ , defined as  $Z$  excluding  $Z_j$  is not restricted to a hyperplane in  $R^{J-1}$ , then the  $c_i, i = 1, \dots, N, \gamma/\sigma_V$  and  $\eta/\sigma_V$  are identified. We augment these conditions by assuming  $V \perp\!\!\!\perp (Z, X_1, \dots, X_N)$  and hence we can identify  $c_i(x_i)$  over the support of  $x_i$ . Using the variation-free condition, if the supports of  $c_i(x_i), c_{i-1}(x_{i-1})$  are unbounded, then by a standard limit argument we can (for each  $Z$ ) set  $\Pr(D_i = 1 \mid Z, X) = 1$  by taking limits.

This produces a standard result where we can identify, in this limit set, the density of  $T \times 1$  vector  $Y_i, g(Y_i)$  associated with choice  $D_i = 1$ . Thus we can use the discussion of Section II, the multifactor case, to identify the factor structure for each choice. If there is sufficient variation in the other  $c(x_j)$ , we can identify  $g(Y_i), j = 1, \dots, N$ . Obviously we cannot identify the

joint density of  $g(Y_1, \dots, Y_N)$  without putting further restrictions on the choice process, as in a Roy model.

## 6 Sampling Algorithm

Let  $S \in \{1, \dots, \bar{S}\}$  denote completed schooling and  $S_T \in \{1, \dots, \bar{S}^*\}$  schooling at test date. The goal is to sample from the posterior distribution of the parameters  $\gamma, \{\delta_j\}_{j=2}^{\bar{S}-1}, \{\alpha_j\}_{j=1}^{J-1}, \{\beta_j, \lambda_j, \sigma_j\}_{j=1}^{\bar{S}^*}$  conditional on observed outcomes  $D, D^T$  and covariates  $Z, W_2, \dots, W_{J-1}$  from a random sample of individuals indexed  $i = 1, \dots, n$ . We impose a noninformative prior on all slope coefficients, i.e.  $p(\gamma, \beta) \propto 1$ . We put proper priors on the variance parameters from the Inverse Gamma family of distributions and on the factor loadings from the Normal distribution family.

We can easily implement a Gibbs sampling algorithm which samples iteratively from the posterior distributions of the parameters conditional on the data and values for the other parameters by augmenting the parameter vector by the latent data  $(V, f, \log \Delta, T^*)$ . The stationary distribution of the Markov chain generated by this algorithm is the joint posterior distribution of the parameters.

The MCMC algorithm is implemented as follows. Given initial starting values for the parameters and  $V, f, \log \Delta, T^*$  for  $m = 1, 2, \dots$  we can update the values of the other parameters and sample from the following conditional distributions (note that we implicitly are conditioning on the data as well as

all other parameters):

1. The conditional posterior distribution of  $V$  is just the product of the individual conditional posterior distributions of  $V_i$  by independence:

$$V_i \sim \text{TN}_{[C_{j-1,i}, C_{ji}]}(Z_i \gamma + \alpha_1 f_i, 1)$$

where  $j = D_i$  and where

$$C_{ji} = \sum_{k=2}^j \Delta_{ki}$$

2. Conditional on  $V$ , the distribution of  $\gamma$  follows from a classical linear regression model with noninformative prior.

$$\gamma \sim N(\hat{\gamma}, \hat{\Omega})$$

where

$$\hat{\gamma} = (Z'Z)^{-1}Z'(V - \alpha_1 f)$$

$$\hat{\Omega} = (Z'Z)^{-1}$$

3. Assuming a normal  $N(\mu_1, \psi_1^2)$  prior the conditional distribution of the factor loading in the choice index is:

$$\alpha_1 \sim N(\widehat{\alpha}_1, \widehat{\Omega}_1)$$

where

$$\begin{aligned}\widehat{\alpha}_1 &= \widehat{\Omega}_1 \left( f'(V - Z\gamma) + \frac{\mu_1}{\psi_1^2} \right) \\ \widehat{\Omega}_1 &= \left( f'f + \frac{1}{\psi_1^2} \right)^{-1}\end{aligned}$$

4. Note above that for those individuals with  $S_i < k$ ,  $\Delta_{ki}$  does not enter the posterior distribution except through its prior distribution, so it can be integrated out (i.e. collected into the normalizing constant). This makes sense, since observations on these people give information only for those transitions that they have attained. Therefore, we only have to sample  $\Delta_{ki}$  for those individuals with  $S_i \geq k$ , i.e., if an individual  $i$  has  $S_i = j$ , then we only have to sample  $\Delta_{ki}$  for  $k = 2, \dots, j$ .

- (a) For  $j = 2, \dots, \bar{S} - 1$ , for  $i$  such that  $S_i = j$ , sample  $\Delta_{ki}$  for  $k = 2, \dots, j$  as follows:

- i. For  $k < j$ , sample  $\Delta_{ki}$  by sampling the truncated normal distribution:

$$\log \Delta_{ki} \sim \text{TN}_{[LB_{ki}^j, UB_{ki}^j]}(W_{ki}\delta_k + \alpha_k f_i, 1),$$

where

$$LB_{ki}^j = \log \left( V_i - \sum_{l=2}^{k-1} \Delta_{li} - \sum_{l=k+1}^j \Delta_{li} \right)$$

$$UB_{ki}^j = \log \left( V_i - \sum_{l=2}^{k-1} \Delta_{li} - \sum_{l=k+1}^{j-1} \Delta_{li} \right)$$

If  $V_i - \sum_{l=2}^{k-1} \Delta_{li} - \sum_{l=k+1}^j \Delta_{li} \leq 0$ , then  $LB_{ki}^j = -\infty$ , i.e. there is no lower bound restriction.

- ii. For  $k = j$ , sample  $\Delta_{ki}$  from the distribution:

$$\log \Delta_{ji} \sim \text{TN}_{[LB_{ji}^j, \infty)}(W_{ki}\delta_k + \alpha_k f_i, 1)$$

where

$$LB_{ji}^j = \log \left( V_i - \sum_{l=2}^{j-1} \Delta_{li} \right)$$

Again, if  $V_i - \sum_{l=2}^{j-1} \Delta_{li} \leq 0$ , then  $LB_{ji}^j = -\infty$ ; in this case the distribution is unrestricted normal.

(b) For  $i$  such that  $S_i = \bar{S}$ , sample  $\Delta_{ki}$  for  $k = 2, \dots, J - 1$  from:

$$\log \Delta_{ki} \sim \text{TN}_{(-\infty, UB_{ki}^J)}(W_{ki}\delta_k + \alpha_k f_i, 1)$$

where

$$UB_{ki}^J = \log \left( V_i - \sum_{l=2}^{k-1} \Delta_{li} - \sum_{l=k+1}^{J-1} \Delta_{li} \right)$$

5. Conditional on  $\Delta_j \equiv \{\Delta_{ji} : S_i \geq j\}$ , the distribution of  $\delta_j$  follows from a classical linear regression model, just as with  $\gamma$ . Let  $\widetilde{W}_j$  be the stacked  $W_{ji}$  for  $i$  such that  $D_i \geq j$ .

For  $j = 2, \dots, \bar{S} - 1$ , sample  $\delta_j$  from the normal distribution:

$$\delta_j \sim \text{N}(\widehat{\delta}_j, \widehat{\Omega}_j)$$

where

$$\begin{aligned}\widehat{\delta}_j &= (\widetilde{W}_j' \widetilde{W}_j)^{-1} \widetilde{W}_j' (\widetilde{\Delta}_j - \alpha_j \widetilde{f}_j) \\ \widehat{\Omega}_j^{(m)} &= \sigma_j^{2, (m-1)} (\widetilde{W}_j' \widetilde{W}_j)^{-1}\end{aligned}$$

6. The conditional distributions of the factor loadings in the cutoffs are as follows, for  $j = 2, \dots, J - 1$ :

$$\alpha_j \sim N(\widehat{\alpha}_j, \widehat{\Omega}_j)$$

where

$$\begin{aligned}\widehat{\alpha}_j &= \widehat{\Omega}_j \left( \widetilde{f}_j' (\widetilde{\Delta}_j - \widetilde{W}_j \delta_j) + \frac{\mu_j}{\psi_j^2} \right) \\ \widehat{\Omega}_j &= \left( \widetilde{f}_j' \widetilde{f}_j + \frac{1}{\psi_j^2} \right)^{-1}\end{aligned}$$

since  $\sigma_j^2 = 1$  and assuming a normal  $N(\mu_j, \psi_j^2)$  prior.

7. For each test equation,  $h = 1, \dots, N_T$ , at each schooling level,  $j =$

$1, \dots, \bar{S}^*$ , we estimate the coefficients on the controls as follows:

$$\beta_{j,h} \sim N\left((X_j'X_j)^{-1}X_j'(T_h^* - \alpha_{j,h}f), \sigma_{j,h}^2(X_j'X_j)^{-1}\right),$$

where only those individuals who have completed schooling level  $j$  at the test date are included.

8. The factor loadings in the test equations are sampled as

$$\lambda_{j,h} \sim N\left(\widehat{\lambda}_{j,h}, \widehat{\Omega}_{j,h}\right)$$

where

$$\begin{aligned}\widehat{\lambda}_{j,h} &= \widehat{\Omega}_{j,h} \left( \frac{f'(T_h^* - X_j\beta_{j,h})}{\sigma_{j,h}^2} + \frac{\mu_{j,h}}{\psi_{j,h}} \right) \\ \widehat{\Omega}_{j,h} &= \left( \frac{f'f}{\sigma_{j,h}^2} + \frac{1}{\psi_{j,h}} \right)^{-1},\end{aligned}$$

using only the individuals who have schooling level  $j$  at the test date and using a normal prior  $N(\mu_{j,h}, \psi_{j,h})$ .

9. Assuming an Inverse Gamma prior  $IG(a_j, b_j)$  and letting  $n^j$  be the

number of individuals in group  $j$  at the test date, we have:

$$\sigma_{j,h}^2 \sim \text{IG} \left( \frac{n^j}{2} + a_j, \frac{(T_h^* - X_j \beta_{j,h} - \lambda_{j,h} f)' (T_h^* - X_j \beta_{j,h} - \lambda_{j,h} f)}{2} + b_j \right)$$

10. The factors  $f$  and the parameters of the factor distribution,

$$p(f) = \sum_{c=1}^{k_m} p_c \text{N}(f | \mu_c, \sigma_c^2)$$

are sampled as follows. Let  $c_i \in \{1, \dots, k_m\}$  denote the mixture component from which  $f_i$  is sampled. Note that  $c_i$  is unobserved. Conditional on  $c_i$  the conditional distribution of  $f_i$  is easily found to be

$$f_i \sim \text{N} \left( \widehat{f}_{ic}, \widehat{\Gamma}_{ic} \right)$$

where

$$\widehat{f}_{ic} = \widehat{\Gamma}_{ic} \left[ \begin{array}{l} \alpha_1 (V_i - Z_i \gamma) + \sum_{k=2}^j (\alpha_k / \sigma_k^2) (\log \Delta_{ki} - W_{ki} \delta_k) + \\ \sum_{h=1}^{N_T} (\lambda_{jT,h} / \sigma_{jT,h}^2) (T_{h,i}^* - X_{jT,i}^h \beta_{jT,h}) + (1 / \sigma_{f,c_i}^2) \mu_{c_i} \end{array} \right]$$

$$\widehat{\Gamma}_{ic} = \left( \sum_{k=2}^j \alpha_k^2 / \sigma_k^2 + \sum_{h=1}^{N_T} \lambda_{jT,h}^2 / \sigma_{jT,h}^2 + 1 / \sigma_{f,c_i}^2 \right)^{-1},$$

and where  $j$  denotes  $S_i = j$  and  $j_T$  denotes  $S_{T,i} = j_T$ .

Conditional on  $f$  the mixture parameters are sampled by the usual trick of first updating the  $c_i$  indicators and then sampling the mixture parameters conditional on the  $c_i$ 's, cf. Robert and Casella (1999). We impose the restriction  $\sum_{c=1}^{kmix} p_c \mu_c = 0$  using the method in Richardson et.al. (2000).

11. The test scores for individuals who hit the ceiling on a test are sampled from truncated normals, i.e.,

$$T_{h,i}^* \sim \text{TN}_{(c_h, \infty)} \left( X'_{s_i,i} \beta_{s_i,h} + \lambda_{s_i,h} f_i, \sigma_{s_i,h}^2 \right).$$

## References

- [1] Anderson, T.W. and H. Rubin. (1956). “Statistical Inference in Factor Analysis,” in J. Neyman, ed. *Proceedings of Third Berkely Symposium on Mathematical Statistics and Probability*, Vol. 5. Berkeley, CA: University of California Press.111-150.
- [2] Angrist, J. and A. Krueger (1991). “Does Compulsory School Attendance Affect Schooling and Earnings?” *Quarterly Journal of Economics*, 106(4): 979-1014.
- [3] Bekker, Paul and J.M.F. ten Berge. (1997). “Generic Global Identification in Factor Models,” *Linear Algebra and Its Applications*, Vol. 264: 255-263.
- [4] Cameron, S. and J. Heckman (1998). “Life Cycle Schooling and Dynamic Selection Bias: Models and Evidence for Five Cohorts of American Males,” *Journal of Political Economy*, 106(2): 262-333.
- [chamber] Chamberlain, G. and Imbens, G. (1996), ”Hierarchical Bayes Model With Many Instrumental Variables”, NBER Technical Working Paper, TO204.

- [5] Ceci, S. J. (1991). "How Much Does Schooling Influence General Intelligence and Its Cognitive Components?: A Reassessment of the Evidence," *Developmental Psychology*, 27: 703-722.
- [6] Chamberlain and Imbens (From Karsten)
- [7] Eckstein, Z. and K. Wolpin (1999). "Why Youths Drop out of High School: The Impact of Preferences, Opportunities, and Abilities," *Econometrica*, 67(6):1295-1339.
- [8] Hanushek, E. (2001). "Black-White Achievement Differences and Governmental Interventions," *American Economic Review*, 91(2): 24-28.
- [9] 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*. Cambridge MA: Cambridge University Press, 156-245.
- [10] Herrnstein, R. and C. Murray (1994) *The Bell Curve*. New York: The Free Press.

- [11] Jencks, C. (1972). *Inequality: a reassessment of the effect of family and schooling in America*. New York: Basic Books.
- [12] Keane, M. and K. Wolpin (1997). "The Career Decisions of Young Men," *Journal of Political Economy*, 105(3), 473-522.
- [13] Kotlarski, I.I. (????). "?????" as cited in Prakasa Rao, B. L. S. *Identifiability in stochastic models: characterization of probability distributions*, Boston: Academic Press.
- [14] Krueger, A. (2001) "Would Smaller Classes Help Close the Black-White Achievement Gap?" Princeton University, Industrial Relations Section, Working Paper #451.
- [15] Ledermann, W., "On the Rank of the Reduced Correlation Matrix in Multiple Factor Analysis," *Psychometrika*, 2, (1937), 85-93.
- [16] Lord, Frederic M. and M. R. Novick. (1968). *Statistical theories of mental test scores*. With contributions by Allan Birnbaum. Reading, Mass. : Addison-Wesley, [1968]

- [17] Moreira, M. (2002) “Tests with Correct Size when Instruments can be Arbitrarily Weak.” Unpublished manuscript, presented at the University of Chicago, September, 2002.
- [18] Mullen, K. J. (2001). “A Generalized Ordered Discrete Choice Model with One-Sided Shocks,” Unpublished manuscript, University of Chicago.
- [19] Neal, D. and W. Johnson (1996). “The Role of Premarket Factors in Black-White Wage Differences,” *Journal of Political Economy*, 104(5): 869-95.
- [20] Rao, B.L.S. Prakasa. (1992). *Identifiability in Stochastic Models*. Boston: Academic Press.
- [21] Richardson, S., L. Leblond, I. Jaussent and P.J. Green. (2000). “Mixture Models in Measurement Error Problems, with Reference to Epidemiological Studies.” Working paper, INSERM, Paris.
- [22] Robert, C.P. and G. Casella. (1999). *Monte Carlo Statistical Methods*, Springer-Verlag, New York.

- [23] Winship, Christopher (2001). “Does Going to College Make You Smarter?” Unpublished manuscript. Harvard University.
- [24] Winship, C. and S. Korenman (1997). “Does Staying in School Make You Smarter? The Effect of Education on IQ in The Bell Curve,” in B. Devlin, S. Fienberg, D. Resnick and K. Roeder, eds., *Intelligence, genes, and success: Scientists respond to The Bell Curve*. New York: Copernicus Press. 215-34.