NBER WORKING PAPER SERIES

ESTIMATING DISTRIBUTIONS OF TREATMENT EFFECTS WITH AN APPLICATION TO THE RETURNS TO SCHOOLING AND MEASUREMENT OF THE EFFECTS OF UNCERTAINTY ON COLLEGE CHOICE

Pedro Carneiro
Karsten T. Hansen
James J. Heckman

Working Paper 9546
http://www.nber.org/papers/w9546

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
March 2003

The views expressed herein are those of the authors and not necessarily those of the National Bureau of Economic Research.

Estimating Distributions of Treatment Effects with an Application to the Returns to Schooling
and Measurement of the Effects of Uncertainty on College Choice
Pedro Carneiro, Karsten T. Hansen and James J. Heckman
NBER Working Paper No. 9546
March 2003
JEL No. C31

ABSTRACT

This paper uses factor models to identify and estimate distributions of counterfactuals. We extend LISREL frameworks to a dynamic treatment effect setting, extending matching to account for unobserved conditioning variables. Using these models, we can identify all pairwise and joint treatment effects. We apply these methods to a model of schooling and determine the intrinsic uncertainty facing agents at the time they make their decisions about enrollment in school. Reducing uncertainty in returns raises college enrollment. We go beyond the "Veil of Ignorance" in evaluating educational policies and determine who benefits and who loses from commonly proposed educational reforms.

Pedro Carneiro
Department of Economics
The University of Chicago
1126 E. 59th Street
Chicago, IL 60637
pmcarnei@midway.uchicago.edu

Karsten T. Hansen
Kellogg School of Management
Northwestern University
2001 Sheridan Rd.
Evanston, IL 60208
karsten-hansen@kellogg.northwestern.edu

James J. Heckman
Department of Economics
The University of Chicago
and The American Bar Foundation
1126 East 59th Street, Chicago, IL 60637
and NBER
jjh@uchicago.edu

1 Introduction

The recent literature on evaluating social programs finds that persons (or firms or institutions) respond to the same policy differently (Heckman, 2001). The distribution of responses is usually summarized by some mean. A variety of means can be defined depending on the conditioning variables used. Different means answer different policy questions. There is no uniquely defined "effect" of a policy.

The research reported here moves beyond means as descriptions of policy outcomes and determines joint counterfactual distributions of outcomes for alternative interventions. From knowledge of the joint distributions of counterfactual outcomes it is possible to determine the proportion of people who benefit or lose from making a particular policy choice (taking or not taking particular treatments), the origin and destination outcomes of those who change states because of policy interventions and the amount of gain (or loss) from various policy choices by persons at different deciles of an initial prepolicy distribution. Our work builds on previous research by Heckman and Smith (1993, 1998) and Heckman, Smith and Clements (1997) that uses experimental data to bound or point-identify joint counterfactual distributions. We extend the analysis of Aakvik, Heckman and Vytlacil (1999, 2003) who use factor models to identify counterfactual distributions to consider indicators for unobservables, implications from choice theory and to exploit the benefits of panel data.

From the joint distribution of counterfactuals, it is possible to generate all mean, median or other quantile gains, to identify all pairwise treatment effects in a multi-outcome setting, and to determine how much of the variability in returns across persons comes from variability in the distributions of the outcome selected and how much comes from variability in opportunity distributions. Using the joint distribution of counterfactuals, it is possible to develop a more nuanced understanding of the distributional impacts of public policies, and to move beyond comparisons of aggregate overall distributions induced by different policies to consider how people in different portions of an initial distribution are affected by public policy. We extend the analysis of DiNardo, Fortin and Lemieux (1996) to consider self-selection as a determinant of aggregate wage and earnings distributions.

Using our methods, we reanalyze the model of Willis and Rosen (1979), who apply the Roy model (1951) to the economics of education. We extend their model to account for uncertainty in the returns to education. We also distinguish between present value income maximizing and utility maximizing evaluations of schooling choices and we estimate the net nonpecuniary benefit of attending college. We use information on the choices of agents to determine how much of the *ex post* heterogeneity in the return to schooling is forecastable at the time agents make their schooling choices. This procedure extends the analysis of Flavin (1981) to a discrete choice setting. This allows us to identify the effect of uncertainty on schooling choices. *Ex ante*, there is a great deal of uncertainty regarding the returns to schooling (in utils or dollars). *Ex post,* 8% of college graduates regret going to college.

The plan of this paper is as follows. Section 2 presents the essential idea underlying the identification strategy used in this paper and how our approach is related to previous work. Section 3 presents a general policy evaluation framework for counterfactual distributions with multiple treatments followed over time. The strategy pursued in this paper is based on using low dimensional factors to generate distributions of potential outcomes. We show how our methods generalize the method of matching by allowing some or all of the variables that generate the conditional independence assumed in matching to be unobserved by the analyst. Section 4 introduces the factor models used in this paper. Section 5 presents proofs of semiparametric identification. Section 6 applies the analysis to extend the Rosen-Willis model of college choice to account for uncertainty and to estimate the information about future earnings available to agents at the time schooling decisions are made. Section 7 reports estimates of the distributions of returns to schooling, the components unforecastable by the agent at the time schooling decisions are made, and the nonpecuniary net benefits from attending college. Section 8 applies our estimates to evaluate a reform of the U.S. educational system. It illustrates the power of our method to lift the commonly invoked Veil of Ignorance and move beyond aggregate distributions of outcomes to understand the consequences of public policies on persons in various parts of the overall distribution. Section 9 concludes. We first provide a brief introduction to the literature to put this paper in context.

2 Estimating Distributions of Counterfactual Outcomes

In order to place the approach used in this paper in the context of an emerging literature on heterogeneous treatment effects, it is helpful to motivate our work by a two outcome, two-treatment cross section model. For simplicity, in this section it is assumed that the outcomes are continuous random variables. The analysis in the rest of this paper is for multiple treatments and multiple outcomes followed over time and the outcomes may be discrete, continuous or mixed discrete-continuous.

The agent can experience one of two possible counterfactual states with associated outcomes (Y_0, Y_1). The states are schooling levels in our empirical analysis. X is a determinant of the counterfactual outcomes (Y_0, Y_1); $S = 1$ if the agent is in state 1; $S = 0$ otherwise. The observed outcome is $Y = SY_1 + (1 - S)Y_0$. There may be an instrument (or set of instruments) Z such that $(Y_0, Y_1) \perp\!\!\!\perp Z \mid X$ and $Pr(S = 1 \mid Z, X)$ depends on Z for all X (*i.e.,* it is a nontrivial function of Z), *i.e., Z* is in the choice probability but not the outcome equation. ($A \perp\!\!\!\perp B \mid C$ means A is independent of B given C). We show below that such a Z is not strictly required in our approach. The standard treatment effect model assumes policies (Z) that affect choices of treatment but not potential outcomes (Y_0, Y_1). General equilibrium effects are ignored.[5]

The goal of our analysis is to recover $F(Y_0, Y_1 \mid X)$. As noted in Heckman (1992), Heckman and Smith (1993, 1998) and Heckman, Smith, and Clements (1997), from this joint distribution it is possible to estimate the proportion of people who benefit (in terms of gross gains) from participation in the program ($Pr(Y_1 > Y_0 \mid X)$), gains to participants at selected levels of the no treatment ($F(Y_1 - Y_0 \mid Y_0 = y_0, X)$) or treatment distribution ($F(Y_1 - Y_0 \mid Y_1 = y_1, X)$), the option value of social programs, and a variety of other questions that can be answered using distributions of potential outcomes including conventional mean treatment effects and quantiles of the gains $(Y_1 - Y_0)$ for those who receive treatment.

The problem of recovering joint distributions arises because we observe Y_0 if $S = 0$ and Y_1 if $S = 1$. Thus we know $F(Y_0 \mid S = 0, X)$, $F(Y_1 \mid S = 1, X)$ but not $F(Y_0 \mid X)$ or $F(Y_1 \mid X)$. In addition, we do not observe the pair (Y_0, Y_1) for anyone. Thus we cannot directly obtain $F(Y_1, Y_0 \mid S, X)$ from the data. Additional information is required to identify the joint distribution.

There are, then, two separate problems. The first is a selection problem. From $F(Y_1 \mid S = 1, X)$ and $F(Y_0 \mid S = 0, X)$, under what conditions can one recover $F(Y_1 \mid X)$ and $F(Y_0 \mid X)$, respectively? The second problem is how to construct the joint distribution $F(Y_0, Y_1 \mid X)$ from the two marginals.

Assuming that the selection problem can be surmounted, classical probability results due to Fréchet (1951) and Hoeffding

(1940) show how to bound $F(Y_1, Y_0 \mid S,\ X)$ using the marginal distributions. In practice these bounds are very wide, and the inferences based on the bounding distributions are often not useful.[6]

The traditional (pre-1985) approach to program evaluation in economics assumed that $F(Y_0, Y_1 \mid X)$ is degenerate because conditional on X, Y_1 and Y_0 are deterministically related:

(1) $$Y_1 \equiv Y_0 + \Delta(X) \quad .$$

This is the "common effect" assumption that postulates that conditional on X, treatment has the same effect on everyone. From the means of $F(Y_0 \mid S = 0,\ X)$ and $F(Y_1 \mid S = 1,\ X)$ corrected for selection, one can identify $E(\Delta(\ X)) = E(Y_1 \mid X) - E(Y_0 \mid X)$. (See Heckman and Robb, 1985; 1986 (reprinted 2000) for a variety of estimators for this case and for discussion of more general cases.) Heckman and Smith (1993, 1998) and Heckman, Smith, and Clements (1997) relax this assumption by assuming perfect ranking across different counterfactual outcome distributions. Assuming absolutely continuous and strictly increasing marginal distributions, they postulate that quantiles are perfectly ranked so $Y_1 = F_{1,X}^{-1}(F_{0,X}(Y_0))$ where $F_{1,X} = F_1(y_1 \mid X)$ and $F_{0,X} = F_0(y_0 \mid X)$. This assumption generates a deterministic relationship which turns out to be the tight upper bound of the Fréchet bounds. An alternative assumption is that people are perfectly inversely ranked so the best in one distribution is the worst in the other: $Y_1 = F_{1,X}^{-1}\left(1 - F_{0,X}(Y_0)\right)$. This is the tight Fréchet lower bound. More generally, one could associate quantiles across distributions more freely. Heckman, Smith and Clements (1997) use Markov transition kernels which stochastically map quantiles of one distribution into quantiles of another. They define a pair of Markov kernels $M(y_1, y_0 \mid X)$ and $\tilde{M}(y_0, y_1 \mid X)$ such that

$$F_1(y_1 \mid X) = \int M(y_1, y_0 \mid X) dF_0(y_0 \mid X)$$

$$F_0(y_0 \mid X) = \int \tilde{M}(y_0, y_1 \mid X) dF_1(y_1 \mid X).$$

Allowing these operators to be degenerate produces a variety of deterministic transformations, including the two previously presented, as special cases of a general mapping. Different $(M, \tilde{M})$ pairs produce different joint distributions.[7] These stochastic or deterministic transformations supply the missing information needed to construct the joint distributions.

A perfect ranking (or perfect inverse ranking) assumption is convenient. It generalizes the perfect-ranking, constant-shift assumptions implicit in the conventional literature. It allows us to apply conditional quantile methods to estimate the distributions of gains.[8] However, it imposes a strong and arbitrary dependence across distributions. Our empirical analysis shows that this assumption is at odds with data on the returns to education.

An alternative approach to constructing joint distributions due to Heckman and Honoré (1990), Heckman (1990) and Heckman and Smith (1998) uses the economics of the model by assuming that

(2) $$S = 1(\mu_s(Z) \geq e_s)$$

where $\mu_s(Z)$ is a mean net utility, $Z \perp\!\!\!\perp e_s$, and "1" is a logical indicator (= 1 if the argument is valid; = 0 otherwise). In addition they assume that

$$\begin{aligned} Y_1 &= \mu_1(X) + U_{1,} \qquad E(U_1) = 0 \\ Y_0 &= \mu_0(X) + U_0, \qquad E(U_0) = 0 \end{aligned}$$

where $(U_1, U_0) \perp\!\!\!\perp (X, Z)$.[9] In the special case where $S = 1(\ Y_1 \geq Y_0)$ (the Roy model), Heckman and Honoré (1990) present conditions on μ_1, μ_0 and X such that $F(U_{1,}U_0)$ and $\mu_1(X)$, $\mu_0(X)$ and hence $F(Y_0,Y_1|X)$ are identified from data on choices (S), characteristics (X) and observed outcomes $Y = SY_1 + (1 - S)Y_{0.}$ Buera (2002) extends their approach to non-separable models with weaker exclusion restrictions.

Heckman (1990) and Heckman and Smith (1998) consider more general decision rules of the form (2) under the assumption that $(Z, X) \perp\!\!\!\perp (U_0, U_1, e_s)$ and the further conditions (i) $\mu_s(Z)$ is a nontrivial function of Z conditional on X and (ii) full support assumptions on $\mu_1(\ X), \mu_0(X)$ and $\mu_s(Z)$. They establish nonparametric identification of $F(U_0, e_s), F(U_1, e_s)$ up to a scale for e_s and $\mu_1(X)$, $\mu_0(X)$ and $\mu_s(Z)$.[10] Hence, under their assumptions, they can identify $F(\ Y_0, S \mid X, Z)$ and $F(\ Y_1,\ S \mid X, Z)$ but *not* the joint distributions $F(\ Y_0,\ Y_1|\ X)$ or $F(\ Y_0,\ Y_1,\ S \mid X,\ Z)$ unless the U_0, U_1, e_s dependence is restricted.

Aakvik, Heckman and Vytlacil (1999, 2003) build on Heckman (1990) and Heckman and Smith (1998) by postulating a factor structure connecting (U_0, U_1, e_s). Our work builds on their analysis so we describe its essential idea. Suppose that the unobservables follow a factor structure:

$$U_0 = \alpha_0\theta + \varepsilon_{0,}\ U_1 = \alpha_1\theta + \varepsilon_1,\ \ e_s = \alpha_s\theta + \varepsilon_s$$

where $\theta \perp\!\!\!\perp (\varepsilon_0, \varepsilon_1, \varepsilon_s)$ and the ε's are mutually independent. In their setup, θ is a scalar. θ can be an unobservable trait like ability or motivation that affects all outcomes. Because the factor loadings, $\alpha_0, \alpha_1, \alpha_s$, may be different, the factors may affect outcomes and choices differently. Recall that one can identify $F(U_0, e_s)$ and $F(U_1, e_s)$ under the conditions specified in Heckman and Smith (1998) and generalized in Theorems 1-3 below. Thus, one can identify $COV(U_0, e_s) = \alpha_0\alpha_s\,\sigma_\theta^2$ and $COV(U_1, e_s) = \alpha_1\alpha_s\sigma_\theta^2$ assuming finite variances and assuming $E(\theta) = 0$, $E(\theta^2) = \sigma_\theta^2$. With some normalizations (e.g., $\sigma_\theta^2 = 1$, $\alpha_s = 1$), under conditions specified in Section 5, we can nonparametrically identify the distribution of θ and the distributions of $\varepsilon_0, \varepsilon_1, \varepsilon_s$ (the last up to scale). With the $\alpha_1, \alpha_0, \alpha_s$, and the distributions of $\theta, \varepsilon_0, \varepsilon_1, \varepsilon_s$ in hand, we can construct the joint distribution $F(Y_0, Y_1 \mid X)$.[11]

This paper builds on this basic idea and extends it to a more general setting. We consider a model with multiple factors, multiple treatments and multiple time periods. Outcome measures may be discrete or continuous. We follow the psychometric literature by adjoining measurement equations to outcome equations to pin down the distribution of θ. With this framework we can estimate all pairwise treatment effects in a multiple outcome setting. We also consider the benefits for identification of having access to imperfect measurements on vector $\boldsymbol{\theta}$ which are observed for all persons independent of their treatment status. This model integrates the *LISREL* framework of Jöreskog (1977) into a model of discrete choice and a model of multiple treatment effects. We develop this model in Section 4 after presenting a more general framework for counterfactuals and treatment effects in a multi-outcome, possibly dynamic setting.

3 Policy Counterfactuals for the Multiple Outcome Case

This section defines policy counterfactuals for the multiple treatment case. For specificity, think of states as schooling levels and different ages as periods in the life cycle. Associated with each state s (schooling level) is a vector of outcomes at age a for person $\omega \in \Omega$ (a set of indices) with elements:

$$Y_{s,a}(\omega) \qquad s = 1, ..., \bar{S},\ a = 1, ..., \bar{A} \tag{3}$$

where there are $\bar{S}$ states and $\bar{A}$ ages. Associated with each person ω is a vector $X(\omega)$ of explanatory variables.

The *ceteris paribus* effect (or individual treatment effect) of a move from state s' at age a'' to state s at age a is

$$\Delta((s,a),(s',a''),\omega) = Y_{s,a}(\omega) - Y_{s',a''}(\omega). \tag{4}$$

Since it is usually not possible to observe the same person in both s and s', analysts often focus on estimating various population level versions of these parameters for different conditioning sets.[12] In this paper, we estimate distributions of potential outcomes and parameters derived from these distributions, including the *Average Treatment Effect*:

$$ATE\left((s,a),(s',a''),x\right) = E(Y_{s,a} - Y_{s',a''} \mid X = x)$$

and the *Marginal Treatment Effect*, the average gain from moving from s' to s for those on the margin of indifference between s and s'. We are interested in determining the joint distributions of the counterfactual distributions of $\Delta((s,a),(s',a''),x)$ for different conditioning sets.

Associated with each treatment or state (schooling choice) is a choice equation associated with a level of lifetime utility: $V_s(\omega),\ s = 1, ..., \bar{S}$. Utilities are assumed to be absolutely continuous. Agents select treatment states (schooling levels) $\tilde{s}$ to maximize utility:

$$\tilde{s} = \underset{s}{\operatorname{argmax}} \left\{V_s(\omega)\right\}_{s=1}^{\bar{S}}. \tag{5}$$

Associated with choices are explanatory variables $Z(\omega)$. A distinctive feature of the econometric approach to program evaluation is that it evaluates policies both in terms of objective outcomes (the $Y_{s,a}(\omega)$) and in terms of subjective outcomes (the utilities of the agents making the choices). Both subjective and objective evaluations are useful in evaluating policy. Choice theory is also used to guide and rationalize specific choices of estimators. It enables us to separate out variability from intrinsic uncertainty, as we demonstrate below.

This framework is sufficiently general to encompass a variety of choice processes including sequential dynamic programming models[13] and ordered choice models,[14] as well as more general unordered choice models. We let $D_s = 1$ if treatment s is selected. Since there are $\bar{S}$ mutually exclusive states, $\sum_{s=1}^{\bar{S}} D_s = 1$.

In this notation, the marginal treatment effect for choices s and s' is

$$\underset{V_s \to V_{s'}}{MTE}(a, \overline{V}_{s,s'}) = E(Y_{s,a} - Y_{s',a} \mid V_s = V_{s'} = \overline{V}_{s,s'} \geq V_j,\ j \neq s, s'). \tag{6}$$

It is the average gain of going from s' to s at age a for persons indifferent between s and s' given that s and s' are the best two choices in the choice set, and that their level of utility is $\overline{V}_{s,s'}$.

Aggregating over choices $s' = 1, ..., \bar{S}$; $s' \neq s$, we may define the marginal treatment effect over all origin states as

$$MTE_s(a, \{\overline{V}_{s,s'}\}_{s'=1, s' \neq s}^{\overline{S}}) = \sum_{\substack{s'=1 \\ s' \neq s}}^{\bar{S}} \underset{V_s \to V_{s'}}{MTE}(a, \overline{V}_{s,s'}) \left(f(V_s, V_{s'} \mid V_s = V_{s'} = \overline{V}_{s,s'} \geq V_j,\ \ j \neq s, s') / \psi(a, \{\overline{V}_{s,s'}\}_{s'=1, s' \neq s}^{\overline{S}}) \right) \tag{7}$$

the weighted average of the pairwise marginal treatment effects from all source states to s (at a given level of utility $\overline{V}_{s,s'}$) with the weights being the density of persons at each relevant margin for specified values of utility where

$$\psi\left(a, \{\overline{V}_{s,s'}\}_{s'=1, s' \neq s}^{\overline{S}}\right) = \sum_{\substack{s'=1 \\ s' \neq s}}^{\overline{S}} f(V_s, V_{s'} \mid V_s = V_{s'} = \overline{V}_{s,s'} \geq V_j,\ j \neq s, s')$$

is a normalizing constant (the population density of people at all margins given a and $\{\overline{V}_{s,s'}\}_{s'=1, s' \neq s}^{\overline{S}}$), assumed to be positive.

We next present a framework for estimating the distributions of the treatment effects and the parameters derived from them, which allows us to estimate the parameters defined in this section as well as other parameters. To simplify notation, we suppress the ω argument in the rest of the paper.

4 Factor Structure Models

The strategy adopted in this paper identifies the distribution of counterfactuals by postulating a low dimensional set of factors $\boldsymbol{\theta}$ so that, conditional on them, and the covariates X and Z, the $Y_{s,a}$ and $V_{s'}$ are jointly independent for all s, s' and a. The distributions of the components of $\boldsymbol{\theta}$ are nonparametrically identified under conditions specified below. With these distributions in hand, it is possible to construct the distribution of counterfactuals. Under the conditions specified in Section 5, it is possible with low dimensional factors to nonparametrically identify the counterfactual distributions and to estimate all of the treatment effects in the literature suitably extended to multidimensional versions.

Throughout this paper we analyze a separable-in-the-errors system. Thus preferences can be described by

$$V_s = \mu_s(Z) - e_s \qquad s = 1, .., \overline{S}. \tag{8}$$

It is conventional to assume that $\mu_s(Z) = Z'\boldsymbol{\beta}_s$ with $s = 1, .., \overline{S}$. Linear approximations to value functions are advocated by Heckman (1981) and Eckstein and Wolpin (1989) and are developed systematically in Geweke, Houser and Keane (2001). Our approach does not require linearity but critically relies on separability between the deterministic portion of the model and the errors e_s. Following Heckman (1981), Cameron and Heckman (1987, 1998), and McFadden (1984), write

$$e_s = \boldsymbol{\alpha}_s' \boldsymbol{\theta} + \varepsilon_s \tag{9}$$

where $\boldsymbol{\theta}$ is a $K \times 1$ vector of mutually independent factors ($\theta_\ell \perp\!\!\!\perp \theta_{\ell'}, \ell \neq \ell'$) and define $\boldsymbol{\varepsilon}^s = (\varepsilon_1, ..., \varepsilon_{\overline{S}})$

$$\boldsymbol{\theta} \perp\!\!\!\perp \boldsymbol{\varepsilon}^s \quad \varepsilon_s \perp\!\!\!\perp \varepsilon_{s'} \quad \forall s, s' = 1, ..., \bar{S} \text{ and } s \neq s' \tag{10}$$

$E(\boldsymbol{\theta}) = \mathbf{0}$; $\quad E(\boldsymbol{\varepsilon}^s) = \mathbf{0}$; $\quad D_k = 1$ if V_k is maximal in $\{V_s(Z)\}_{s=1}^{\overline{S}}$.[15]

Potential outcomes at age a, $Y_{s,a}^*$, are stochastically dependent among each other and the choices only through their dependence on the observables X,Z and the factors $\boldsymbol{\theta}$:

$$Y_{s,a}^* = \mu_{s,a}(X) + \boldsymbol{\alpha}_{s,a}' \boldsymbol{\theta} + \varepsilon_{s,a} \tag{11}$$

where $E(\varepsilon_{s,a}) = 0$.

Potential outcomes are separable in observables and unobservables. A linear-in-parameters version writes $\mu_{s,a}(X) = X' \boldsymbol{\beta}_{s,a}$. Define $\boldsymbol{\varepsilon}^Y$=$(\varepsilon_{1,1}, ..., \varepsilon_{1,\overline{A}}, ..., \varepsilon_{s,1}, ..., \varepsilon_{s,\overline{A}}, ..., \varepsilon_{\overline{S},\overline{A}})$

$$\boldsymbol{\theta} \perp\!\!\!\perp \boldsymbol{\varepsilon}^Y \tag{12}$$

$$\varepsilon_{s,a} \perp\!\!\!\perp \varepsilon_{s',a''}; \quad \forall\ s \neq s'; \quad \forall\ a, a'' \tag{13}$$

and

$$\varepsilon_{s,a} \perp\!\!\!\perp \varepsilon_{s'}; \quad \forall\ s',\ s = 1, ..., \bar{S}; \quad a = 1, ..., \bar{A}. \tag{14}$$

$$(Z, X) \perp\!\!\!\perp (\boldsymbol{\theta}, \boldsymbol{\varepsilon}^Y, \boldsymbol{\varepsilon}^s) \tag{15}$$

The $Y_{s,a}^*$ may be vector valued.

When the outcome is continuous, the observed value corresponds to the latent variable ($Y_{s,a} = Y_{s,a}^*$). When the outcome is discrete (*e.g.,* employment status), we interpret $Y_{s,a}^*$ in (11) as a latent variable. In that case, $Y_{s,a}$ is an indicator function $Y_{s,a} = 1(Y_{s,a}^* \geq 0)$. Tobit and other censored cases can be accommodated. Other mixed discrete-continuous cases can be handled in a conventional fashion.[16]

One motivation for the factor representation is that agents may observe components of $\boldsymbol{\theta}$ (or variables that span those components) and act on them (*e.g.,* choose schooling levels), while the econometrician does not observe $\boldsymbol{\theta}$. Below, we present methods for testing whether agents observe some or all components of $\boldsymbol{\theta}$. Conditional on $\boldsymbol{\theta}$ and X, the potential outcomes are independent. If (12)-(15) accurately describe the data generating process, we obtain the conditional independence assumptions used in matching (see, *e.g.,* Cochrane and Rubin, 1973; Rosenbaum and Rubin, 1983).

In matching it is assumed that $Y_{s,a} \perp\!\!\!\perp D_s \mid X, Z, \boldsymbol{\theta}$ for all s.[17] From this assumption, we can identify ATE from the right hand side of the following expression for continuous observed outcomes, which can be constructed if $\boldsymbol{\theta}$ is observable:

$$E(Y_{s,a} - Y_{s',a} \mid X, Z, \boldsymbol{\theta}) = E(Y_{s,a} \mid X, Z, \boldsymbol{\theta}, D_s = 1) - E(Y_{s',a} \mid X, Z, \boldsymbol{\theta}, D_{s'} = 1).$$

In this case treatment on the treated, ATE and MTE are the same parameter conditional on $\boldsymbol{\theta}$, X and Z (Heckman, 2001; Aakvik, Heckman and Vytlacil, 2003). Our framework differs from matching by allowing the factors that generate the conditional independence that underlies matching to be unobserved by the analyst. In this sense, our approach is more robust than matching. The price for this robustness is the assumed independence between $\boldsymbol{\theta}$ and (X, Z).

Factor structure models are notorious for being identified by arbitrary normalization and exclusion restrictions. To reduce this arbitrariness and render greater interpretability to estimates obtained from our model, we adjoin a measurement system to choice equations (8) and outcome system (11). Various measurements can be interpreted as indicators of specific factors (*e.g.,* test scores may proxy ability). Having measurements on the factors also facilitates identifiability under weaker assumptions as we demonstrate in Section 5. However, measurements are *not* strictly required for identification in our model. Outcome, measurement, and choice equations are interchangeable sources of identification in a sense that we make precise in Section 5.

Consider a system of L measurements on the K factors, initially assumed to be for continuous outcome measures:

$$\begin{aligned} M_1 &= \mu_1(X) + \beta_{11}\theta_1 + ... + \beta_{1K}\theta_K + \varepsilon_1^M \\ &\vdots \\ M_L &= \mu_L(X) + \beta_{L1}\theta_1 + ... + \beta_{LK}\theta_K + \varepsilon_L^M \end{aligned} \tag{16}$$

$\boldsymbol{\varepsilon}^M = (\varepsilon_1^M, ..., \varepsilon_L^M)$, $E(\boldsymbol{\varepsilon}^M) = \mathbf{0}$ and where we assume $\boldsymbol{\theta} \perp\!\!\!\perp \boldsymbol{\varepsilon}^M$, $\boldsymbol{\theta} \perp\!\!\!\perp \boldsymbol{\varepsilon}^s$, $\boldsymbol{\varepsilon}^s \perp\!\!\!\perp (\boldsymbol{\varepsilon}^M, \boldsymbol{\varepsilon}^Y)$, $\varepsilon_i^M \perp\!\!\!\perp \varepsilon_j^M$ $\forall i \neq j$, and $i, j = 1, \ldots, L$. For interpretability, we assume $\theta_i \perp\!\!\!\perp \theta_j$, $\forall i \neq j$, $i, j = 1, ..., K$. We develop the case with discrete measurements on latent continuous variables in Section 5. One can think of the outcome measures as an s-dependent measurement system. The measures (16) are the same across all s.

Measurement system (16) allows for fallible measures of outcomes. Thus in our schooling choice analysis we are not committed to the infallibility of test scores as measurements of ability. Measurement system (16) allows us to proxy unobservables accounting for measurement error and hence enables us to improve on the proxy procedure of Olley and Pakes (1996) which assumes no measurement error.

Choice Equations

Our analysis applies to both ordered discrete choice models and unordered choice models as analyzed by Cameron and

Heckman (1998) and Hansen, Heckman and Mullen (2001). In this paper, we focus attention on a new ordered choice model. Other choice models can easily be accommodated in our framework and richer models are a source of additional identifying information.[18]

For an ordered discrete choice model, let utility index I be written as

$$I = \varphi(Z) + \varepsilon_W, \quad \varepsilon_W = \boldsymbol{\gamma}'\boldsymbol{\theta} + \varepsilon_I, \; \sigma_W^2 = \boldsymbol{\gamma}' \textstyle\sum_{\boldsymbol{\theta}} \boldsymbol{\gamma} + \sigma_I^2 \tag{17}$$

where $E(\varepsilon_I^2) = \sigma_I^2$, and $\sum_{\boldsymbol{\theta}}$ is the covariance matrix of $\boldsymbol{\theta}$. A linear-in-parameters version which is the one developed in this paper writes $\varphi(Z) = Z\eta$. Choices are generated by index $\varphi(Z)$ falling in various intervals.

$$\begin{aligned} D_1 &= 1 \text{ if } -\infty < I \leq c_1 \\ D_s &= 1 \Leftrightarrow c_{s-1} < I \leq c_s \qquad s = 2, ..., \bar{S} - 1 \\ D_{\overline{S}} &= 1 \text{ if } c_{\overline{S}-1} < I < \infty \end{aligned} \tag{18}$$

where $c_0 = -\infty$. It is required that $c_s \geq c_{s-1}$ for all $s \geq 2$. This is a special case of random utility model (8) in which states are ordered and pairwise contrasts possess a special structure.[19]

We can parameterize the c_s to be functions of state-specific regressors, *e.g.*, $c_s = Q_s\rho_s$ where we restrict $c_s \geq c_{s-1}$. We could also follow a suggestion in Heckman, LaLonde and Smith (1999) and incorporate one sided shocks ν_s and work with stochastic thresholds $\tilde{c}_s$ in place of $c_s : \tilde{c}_s = c_s + \nu_s, \;\; s = 1, ..., \overline{S} - 1$ where $\nu_s \geq \nu_{s-1}$ and $\nu_s \geq 0$.[20]

Conditioning on $Q_s = q_s, s = 1, ..., \bar{S}$, and assuming that the Support$(Z \mid Q_s = q_s, s = 1, ..., \bar{S}) =$ Support(ε_W), we can apply the conditions presented in Cameron and Heckman (1998) to identify the distribution of F_{ε_W}, η, $c_1, ..., c_{\overline{S}-1}$ up to scale. We can nonparametrically identify $c_s(Q_s)$ over the support of Q_s under conditions specified in Theorem 2 below. Unlike the case of the more general unordered discrete choice model (see Elrod and Keane, 1995; Ben Akiva et al., 2001), without further restrictions on the distribution of ε_W, we cannot identify the factors generating ε_W using only choice data. Hansen, Heckman and Mullen (2003) present an analysis parallel to the one given here for a multinomial probit model. In that model, the distributions of factors can be identified from choice data.

4.1 Models for Factors

Factor models are notorious for being identified through arbitrary assumptions about how factors enter in different equations. This led to their disuse after their introduction into economics by Jöreskog and Goldberger (1975), Goldberger (1972), Chamberlain and Griliches (1975) and Chamberlain (1977a, b).

The essential identification problem in factor analysis is clearly stated by Anderson and Rubin (1956). If there are L measurements on K mutually independent factors arrayed in a vector $\boldsymbol{\theta}$, we may write outcomes G in terms of latent variables $\boldsymbol{\theta}$ as

$$G = \boldsymbol{\mu} + \Lambda\boldsymbol{\theta} + \boldsymbol{\varepsilon} \tag{19}$$

where G is $L \times 1$, $\boldsymbol{\theta} \perp\!\!\!\perp \boldsymbol{\varepsilon}$, $\boldsymbol{\mu}$ is an $L \times 1$ vector of means, which may depend on X, $\boldsymbol{\theta}$ is $K \times 1$, $\boldsymbol{\varepsilon}$ is $L \times 1$ and $\boldsymbol{\Lambda}$ is $L \times K$. $\varepsilon_i \perp\!\!\!\perp \varepsilon_j$, $i, j = 1, ., L, i \neq j$. At this point, $\boldsymbol{\varepsilon}$ is a general notation which will be linked to specific ε's in Section 5. Even if $\theta_i \perp\!\!\!\perp \theta_j$, $i \neq j, i, j = 1.., K$, the model is underidentified. As we shall see, the G in this paper is a more general system than the system based solely on measurements invariant across states M so we distinguish (16) and (19). It will include M as well as state dependent outcomes $(Y^*_{s,a})$ and the indices generating choice equations.

Using only the information in the covariance matrices, as is common in factor analysis,

$$COV(G) = \Lambda \Sigma_{\boldsymbol{\theta}} \Lambda' + D_{\boldsymbol{\varepsilon}} \tag{20}$$

where $\Sigma_{\boldsymbol{\theta}}$ is a diagonal matrix of the variances of the factors, and $D_{\boldsymbol{\varepsilon}}$ is a diagonal matrix of the "uniqueness" variances. We observe G but not $\boldsymbol{\theta}$ or $\boldsymbol{\varepsilon}$, and we seek to identify Λ, $\Sigma_{\boldsymbol{\theta}}$ and $D_{\boldsymbol{\varepsilon}}$. Without some restrictions, this is clearly an impossible task. Conventional factor-analytic models make assumptions to identify parameters. The restriction that the components of $\boldsymbol{\theta}$ are independent is one restriction that we have already made, but it is not enough. The diagonals of $COV\,(G)$ combine elements of D_ε with parameters from the rest of the model. Once those other parameters are determined, the diagonals identify D_ε. Accordingly, we can only rely on the $\frac{L(L-1)}{2}$ non-diagonal elements to identify the K variances (assuming $\theta_i \perp\!\!\!\perp \theta_j$, $\forall i \neq j$), and the $L \times K$ factor loadings. Since the scale of each θ_i is arbitrary, one factor loading devoted to each factor is normalized to unity to set the scale. Accordingly, we require that

$$\underbrace{\frac{L\,(L-1)}{2}}_{\text{Number of off-diagonal covariance elements}} \geq \underbrace{(L \times K - K)}_{\text{Number of unrestricted } \Lambda} + \underbrace{K}_{\text{Variances of } \theta}$$

so

$$L \geq 2K + 1$$

is a necessary condition for identification.

The strategy pursued in this paper is transparent and assumes that there are two or more elements of G devoted *exclusively* to factor θ_1, and at least three elements of G that are generated by factor θ_1, two or more other elements of G devoted only to factors θ_1 and θ_2, with at least three elements of G that depend on θ_1 and θ_2, and so forth. This strategy is motivated by our access to psychometric and longitudinal data. Test scores may only proxy ability (θ_1). Other measurements may proxy only (θ_1, θ_2). Measurements on earnings from panel data may proxy $(\theta_1, \theta_2, \theta_3)$, etc.

Order G under this assumption so that we get the following pattern for Λ (we assume that the displayed λ_{ij} are not zero):

$$\Lambda = \left(\begin{array}{ccccccccc} 1 & 0 & 0 & 0 & \vdots & ... & ... & 0 \\ \lambda_{21} & 0 & 0 & 0 & \vdots & ... & ... & 0 \\ \lambda_{31} & 1 & 0 & 0 & \vdots & ... & ... & 0 \\ \lambda_{41} & \lambda_{42} & 0 & 0 & \vdots & ... & ... & 0 \\ \lambda_{51} & \lambda_{52} & 1 & 0 & \vdots & ... & ... & 0 \\ \lambda_{61} & \lambda_{62} & \lambda_{63} & 0 & \vdots & ... & ... & 0 \\ \lambda_{71} & \lambda_{72} & \lambda_{73} & 1 & \vdots & 0 & ... & 0 \\ \lambda_{81} & \lambda_{82} & \lambda_{83} & \lambda_{84} & \vdots & 0 & ... & 0 \\ ... & ... & ... & ... & \vdots & ... & ... & ... \\ \lambda_{L,1} & \lambda_{L,2} & \lambda_{L,3} & ... & \vdots & ... & ... & \lambda_{L,K} \end{array} \right). \tag{21}$$

Assuming nonzero covariances

$$COV(g_j, g_l) = \lambda_{j1}\lambda_{l1}\sigma^2_{\theta_1}, \quad l = 1, 2; \quad j = 1, ..., L; \quad j \neq l.$$

In particular

$$\begin{aligned} COV(g_1, g_\ell) &= \lambda_{\ell 1}\sigma^2_{\theta_1} \\ COV(g_2, g_\ell) &= \lambda_{\ell 1}\lambda_{21}\sigma^2_{\theta_1}. \end{aligned}$$

Assuming $\lambda_{\ell 1} \neq 0$, we obtain

$$\frac{COV(g_2, g_\ell)}{COV(g_1, g_\ell)} = \lambda_{21}.$$

Hence, from $COV(g_1, g_2) = \lambda_{21}\sigma^2_{\theta_1}$, we obtain $\sigma^2_{\theta_1}$, and hence $\lambda_{\ell 1}$, $\ell = 1, \ldots, L$. We can proceed to the next set of two measurements and identify

$$COV(g_l, g_j) = \lambda_{l1}\lambda_{j1}\sigma^2_{\theta_1} + \lambda_{l2}\lambda_{j2}\sigma^2_{\theta_2}, \quad l = 3, 4; \quad j \geq 3; \quad j \neq l.$$

Since we know the first term on the right hand side by the previous argument, we can proceed using $COV(g_l, g_j) - \lambda_{l1}\lambda_{j1}\sigma^2_{\theta_1}$ and identify the $\lambda_{j2}, j = 1, ..., L$ using the previous line of reasoning (some of these elements are fixed to zero). Proceeding in this fashion, we can identify Λ and $\Sigma_{\boldsymbol{\theta}}$ subject to diagonal normalizations. This argument works for all but the system for the K^{th} and final factor. Observe that for all of the preceding factors there are at least three measurements that depend on $\theta_j, j = 1, \ldots, K - 1$, although only two of the measurements need to depend solely on $\theta_{1,\ldots,}\theta_{K-1}$. To obtain the necessary three measurements for the K^{th} and final factor, we require that there be at least three outcomes with measurements that depend on $\theta_1, \ldots, \theta_K$.

Knowing Λ and $\Sigma_{\boldsymbol{\theta}}$, we can identify D_{ε}. Use of dedicated measurement systems for specific factors and panel data helps to eliminate much of the arbitrariness that plagued factor analysis in its 1970's introduction in economics. While many other restrictions on the model are possible, the one we adopt has the advantage of simplicity and interpretability in many contexts.[21]

Our analysis uses a version of (19), coupled with the exclusion restrictions exemplified in (21), to identify the joint distributions of counterfactuals. We extend conventional factor analysis in three ways. First, following Heckman (1981) and Muthen (1984), we allow the G to include latent index functions like I (associated with the choice equations) or like $Y^*_{s,a}$ as well as their manifestations (the random variables they generate). Thus the G may include discrete or censored random variables generated by latent random variables. We can identify components associated solely with the discrete case only up to unknown scale factors – the familiar indeterminacy in discrete choice analysis. Choice indices, measurements and state contingent outcomes are all informative on $\boldsymbol{\theta}$. The factor analysis in this paper is conducted on the latent continuous variables that generate the manifest outcomes. Second, we extend factor analysis to a case with counterfactuals where certain variables are only observed if state s is observed. This extension enables us to identify the full joint distribution of counterfactuals. Third, we prove nonparametric identification of the distributions of $\boldsymbol{\theta}$ and $\boldsymbol{\varepsilon}$, and do not rely on any normality assumptions.

5 Identification of Semiparametric Factor Models with Discrete Choices and Discrete and Continuous Outcomes

In order to establish identification, we need to be clear about the raw data with which we are working. For each set of s-contingent potential outcomes, there is a system like (19): $\widetilde{G}_s = (M, Y_s, D_s)$ where Y_s is a vector of state contingent outcomes. Outcome variables in $\widetilde{G}_s$ are of two types: (a) continuous variables and (b) discrete or censored random variables, including binary strings associated with durations (*e.g.*, unemployment). When the random variables are discrete or censored, we work with the latent variables generating them. We array the continuous portions of $\widetilde{G}_s$ and the index functions generating the discrete portions into G_s.

Let M^c denote the continuous measurements, and let Y^c_s be the continuous counterfactual outcomes. Let M^d be the discrete components of M, while Y^d_s are the discrete components of Y_s. Table 1 defines the variables used in our analysis.

Under separability the continuous variables can be written as

$$\begin{aligned} M^c &= \boldsymbol{\mu}^c_m(X) + U^c_m \\ Y^c_s &= \boldsymbol{\mu}^c_s(X) \;\; + U^c_s. \end{aligned}$$

Associated with the discrete variables are latent continuous variables

$$\begin{aligned} M^{*d} &= \boldsymbol{\mu}^d_m(X) + U^d_m \\ Y^{*d}_s &= \boldsymbol{\mu}^d_s(X) \;\; + U^d_s \end{aligned}$$

where U_m^d, U_s^d are assumed to be continuous.[22] The indicator variable is generated by latent variable I as defined in (17).

The data used for the factor analysis are $G_s = (M^c,\ M^{*d},\ Y_s^c,\ Y_s^{*d},\ I)$. For simplicity, in this paper we assume that the "discrete" variables are in fact binary valued. Extensions to censored random variables and to binary strings are straightforward and are developed in a later paper. We observe $\widetilde{G}_s$ when $D_s = 1$. For each s, we have a system of outcome variables. While the outcomes are s-dependent, the measurements are observed independently of the value assumed by D_s.

The distinction between measurements (M) whose values do not depend on the value assumed by D_s, and the state contingent outcomes Y_s that depend on the state s that is observed, is essential. There is no selection bias in observing M but in general there is selection bias in observing $Y = \sum_{s=1}^{\overline{S}} D_s Y_s$.

M, Y, and $D_s, s = 1, .., \overline{S}$ all contain information on $\boldsymbol{\theta}$. The information from M is easier to access, and traditional factor analysis is based on such measurements. Nonetheless, the identification of counterfactual states does not require M. If M is available, however, the interpretation of $\boldsymbol{\theta}$ is more transparent.

Before turning to our factor analysis, we first establish conditions under which we can identify the joint distribution of $M^c, M^{*d}, Y_s^c, Y_s^{*d}, I$, which constitute the data for the factor analysis. To understand the basic ideas, we break this task into three parts: (a) identification of the joint distribution of $\left(M^c, M^{*d}\right)$; (b) identification of the parameters in choice system (17) and (18) and (c) identification of the full joint distribution of $\left(M^c, M^{*d}, Y_s^c, Y_s^{*d}, I\right)$. This full distribution is subsequently factor analyzed.

We assume that

(A-1) $\left(U_m^c, U_m^d, U_s^c, U_s^d, \varepsilon_W\right)$ *have distribution functions that are absolutely continuous with respect to Lebesgue measure with means zero*[23] *with support* $\mathbb{U}_m^c \times \mathbb{U}_m^d \times \mathbb{U}_s^c \times \mathbb{U}_s^d \times \mathbb{E}_W$ *with upper and lower limits being* $\bar{U}_m^c, \bar{U}_m^d, \bar{U}_s^c, \bar{U}_s^d, \overline{\varepsilon}_W$ *and* $\underline{U}_m^c, \underline{U}_m^d, \underline{U}_s^c, \underline{U}_s^d, \underline{\varepsilon}_W$, *respectively, which may be bounded or infinite. Thus the joint system is measurably separable (variation free).*[24] *We assume finite variances.*[25] *The cumulative distribution function of* ε_W *is assumed to be strictly increasing over its full support* $(\underline{\varepsilon}_W, \overline{\varepsilon}_W)$.

(A-2) $(X, Z, Q) \perp\!\!\!\perp (U, \varepsilon_W)$ *where* $U = (U_m^c,\ U_m^d,\ U_s^c,\ U_s^d)$, *where* Q *is a vector of state-specific regressors* $Q = (Q_1, \ldots, Q_{\overline{S}})$.

We denote by "~" normalized values where the normalizations in our context are usually standard deviations of latent index errors. We first consider identification of the joint distribution of M. Our results are contained in Theorem 1.

Theorem 1 *From data on* $F(M \mid X)$, *one can identify the joint distribution of* $\left(U_m^c, U_m^d\right)$ *(the latter component only up to scale), the function* $\boldsymbol{\mu}_m^d(X)$ *is identified and* $\boldsymbol{\mu}_m^c(X)$ *is identified over the support of* X *(up to scale) provided that the following assumptions, in addition to the relevant components of (A-1) and (A-2), are invoked.*

(A-3) *Order the discrete measurement components to be first. Suppose that there are* $N_{m,d}$ *discrete components, followed by* $N_{m,c}$ *continuous components. Assume Support*$\left(\mu_{1,m}^d(X), \ldots, \mu_{N_{m,d},m}^d(X)\right) \supseteq$ *Support* $\left(U_{1,m}^d, ..., U_{N_m,m}^d\right)$.

Conditions (A-1) and (A-3) imply that $\left(\mu_{1,m}^d(X), \ldots, \mu_{N_{m,d},m}^d(X)\right)$ is measurably separable (variation free) in all of its coordinates when "$\supseteq$" is replaced by "$=$."

(A-4) *For each* $l = 1, ..., N_{m,d}$ $\ \mu_{l,m}^d(X) = X\beta_{l,m}^d$.

(A-5) *The X lives in a subset of $\mathbb{R}^{N_X}$. There exists no linear proper subspace of $\mathbb{R}^{N_X}$ having probability* 1 *under F_X, the distribution function of X.*

Proof: See Appendix A.

Condition (A-4) is conventional (See Cosslett, 1983, or Manski, 1988). Weaker conditions are available using the analysis of Matzkin (1992,1993). Support condition (A-3) appears in Cameron and Heckman (1998) and Aakvik, Heckman and Vytlacil (1999). The easiest way to satisfy it is to have exclusions: one continuous component in $\mu^d_{l,m}(X)$ that is not an argument in the others. But that is only a sufficient condition. Even without exclusion, this condition can be satisfied if there are enough continuous regressors in X and the $\mu^d_{l,m}(X)$ have a full rank Jacobian - with respect to the derivatives of the continuous (X) variables. Intuitively, if the rank condition is satisfied, we can hold $\mu^d_{l,m}(X)$ at $\bar{\mu}^d_{l,m}$ and vary the other arguments. Formally, this rank condition requires that if we array the coefficients of the continuous variables coefficients of $\beta^d_{l,m}, \widetilde{\beta}^d_{l,m}$, into a $N_{m,d}$ by N_X matrix, where N_X is the number of continuous components of X, that *Rank* $\left\{\widetilde{\beta}^d_{l,m}\right\}_{l=1}^{N_{m,d}} \geq N_{m,d}$. This requires $N_{m,d}$ continuous variables. It also requires that the coefficients are linearly independent. If the number of continuous components is $N_X < N_{m,d}$, we can only identify N_X components of the distribution of U^d_m. We can trace out the distribution of the latent variables even if the X are not of full rank, so (A-5) is not strictly required. Observe that we can identify the joint distribution of the U_m even if all components of β are not identified because of a failure of a rank condition. See Cameron and Heckman (1998), Aakvik, Heckman and Vytlacil (1999) or Hansen, Heckman and Mullen (2003) for more discussion of this case of identification without conventional exclusion restrictions.

We next turn to identification of the generalized ordered discrete choice model (17). This extends the proof in Cameron and Heckman (1998) by parameterizing the cut points. A more general version of this model appears in Hansen, Heckman and Mullen (2001).

Theorem 2 *For the relevant subsets of the conditions (A-1), and (A-2) (specifically, assuming absolute continuity of the distribution of ε_W with respect to Lebesgue measure and $\varepsilon_W \perp\!\!\!\perp (Z, Q)$), and the additional assumptions:*

(A-6) $c_s(Q_s) = Q_s\eta_s, s = 1, ..., \overline{S}, \varphi(Z) = Z'\boldsymbol{\beta}$

(A-7) (Q_1, Z) *is full rank (there is no proper subspace of the support* (Q_1, Z) *with probability* 1). *The Z contains no intercept.*

(A-8) Q_s for $s = 2, \ldots, \overline{S}$ is *full rank (there is no proper subspace of* $\left(\mathbb{R}^{Q_s}\right)$ *with probability* 1*).*

(A-9) *Support* $(\mathbf{c}(Q_1) - \varphi(Z)) \supseteq$ *Support* (ε_W)

Then the distribution function F_{ε_W} is known up to a scale normalization on ε_W and $c_s(Q_s), s = 1, ...\bar{s}$, and *$\varphi(Z)$ are identified up to a scale normalization.*

Proof: See Appendix A.

Our choice system can be made nonparametric using the type of restrictions introduced in Matzkin, although we eschew that generality here. Matzkin and Lewbel (2002) weaken (A-6) generalizing the analysis of Matzkin (1992) assuming that the c_s are constants.

We next turn to the identification of the joint system $(M^c, M^{*d}, Y_s^c, Y_s^{*d}, I)$. The data for each choice system (including the data on choice probabilities) generate the left hand side

$$\Pr\left(M^c \leq m^c, M^{*d} \leq 0, Y_s^c \leq y_s^c, Y_s^{*d} \leq 0 | D_s = 1, X, Z, Q_s, Q_{s-1}\right) \Pr(D_s = 1|Z, Q_s, Q_{s-1}) \tag{22}$$

$$= \int_{\underline{U}_c}^{m^c - \boldsymbol{\mu}_m^c(X)} \int_{\underline{\widetilde{U}}_m^d}^{-\widetilde{\boldsymbol{\mu}}_m^d(X)} \int_{\underline{U}_s^c}^{y_s^c - \boldsymbol{\mu}_s^c(X)} \int_{\underline{\widetilde{U}}^d}^{-\widetilde{\boldsymbol{\mu}}^d(X)} \int_{\frac{c_{s-1}(Q_{s-1}) - \varphi(Z)}{\sigma_W}}^{\frac{c_s(Q_s) - \varphi(Z)}{\sigma_W}} f\left(U_{m,}^c \widetilde{U}_{m,}^d U_{s,}^c \widetilde{U}_{s,}^d\ \widetilde{\varepsilon}_W\right) dU_m^c d\widetilde{U}_m^d dU_s^c d\widetilde{U}_s^d d\widetilde{\varepsilon}_W.$$

From Theorem 1 we know $\boldsymbol{\mu}_m^c(X)$ $(= X\beta_m^c)$ and $\widetilde{\boldsymbol{\mu}}_m^d(X)$ $(= X\widetilde{\beta}_m^d)$ and the joint distribution of $(U_{m,}^c \widetilde{U}_m^d)$. From Theorem 2, we know $\frac{c_s(Q_s) - \varphi(Z)}{\sigma_W} = \frac{Q_s \eta_s - Z'\beta}{\sigma_W}$, $s = 1, ..., \overline{S}$ and the coefficients η_s, β and the distribution $F_{\widetilde{\varepsilon}_W}$. Notice that $c_s(Q_s) \geq c_{s-1}(Q_{s-1})$ is a requirement of the ordered choice model. We maintain the following assumptions:

(A-10) *Support* $\left(-\widetilde{\boldsymbol{\mu}}_m^d(X), -\widetilde{\boldsymbol{\mu}}_s^d(X), \left(\frac{c_s(Q_s) - \varphi(Z)}{\sigma_W} - \frac{c_{s-1}(Q_{s-1}) - \varphi(Z)}{\sigma_W}\right)\right) \supseteq$*Support*$(U_m^d, U_s^d, \widetilde{\varepsilon}_W) = (\mathbb{U}_m^d \times \mathbb{U}_s^d \times \widetilde{\mathbb{E}}_W)$.

(A-11) *There is no proper linear subspace of* (X, Z, Q_s, Q_{s-1}) *with probability one so the model is full rank.*

As a consequence of (A-6) and (A-10) we can find values of $Q_s, Q_{s-1}, \bar{Q}_s, \underline{Q}_{s-1}$ respectively so that

$$\lim_{\substack{Q_s \to \bar{Q}_s \\ Q_{s-1} \to \underline{Q}_{s-1}}} \Pr\left(D_s = 1|Z, Q_s, Q_{s-1}\right) = 1.$$

In these limit sets (which may depend on Z), under the stated conditions (A-1) – (A-11), we can identify the joint distribution of $(M^c, M^{*d}, Y_s^c, Y_s^{*d})$, $s = 1, \ldots, \overline{S}$ using an argument parallel to the one used to prove Theorem 1. These limit sets produce $\overline{S}$ different joint distributions (corresponding to each value of s) but do not generate joint distributions *across* the s (*i.e.,* the joint distribution of $M^c, M^{*d}, Y_s^c, Y_s^{*d}$ across s values). However, M is common across these systems. Using the dependence of M and $Y_s, s = 1, \ldots, \overline{S}$ on a common $\boldsymbol{\theta}$ we can sometimes identify the joint distribution. See Carneiro, Hansen and Heckman (2001) for an example. Thus with a measurement system M we do not strictly require information on the choice index I to identify the model.

Following an argument of Heckman (1990), Heckman and Honoré (1990) and Heckman and Smith (1998), we can identify $\boldsymbol{\mu}_s^c(X)$ up to an additive constant without passing to the limit set where $\Pr(D_s = 1|Z, Q_s, Q_{s-1}) = 1$. This is not possible for the identification of $\widetilde{\boldsymbol{\mu}}_s^d(X)$ because there is no counterpart to the variation in y_s^c for the discrete component. This is the content of the following theorem which combines the key ideas of Theorems 1 and 2 to produce an identification theorem for the general case.

Theorem 3 *Under assumptions (A-1), (A-2), (A-4), (A-6), (A-7),(A-8),(A-9),(A-10) and (A-11),* $\boldsymbol{\mu}_m^c(X), \boldsymbol{\mu}_s^c(X), \widetilde{\boldsymbol{\mu}}_m^d(X)$, $\widetilde{\boldsymbol{\mu}}_s^d(X), \widetilde{\varphi}(Z), c_s(Q_s)$ $s = 1, ..., \overline{S} - 1$ *are identified as is the joint distribution* $F(U_m^c, \tilde{U}_m^d, U_s^c, \tilde{U}_s^d, \widetilde{\varepsilon}_W)$.

Proof: See Appendix A.

As noted in the discussion following Theorem 1, without standard exclusion restrictions we may only be able to identify subcomponents of the joint distribution if $N_X < N_{m,d}$ where N_X is the number of continuous regressors. Note that the $\mu_{s,l}^c$, $\widetilde{\mu}_{s,l}^d$ may only be defined over their supports. Under an additional rank or variation-free condition on the regressors we recover these functions everywhere over the support of X.

5.1 Factor Analysis

The thrust of Theorems 1-3 is that under the stated conditions we know the joint distributions of $(U_s, U_m, \tilde{\varepsilon}_W)$ $s = 1, ..., \overline{S}$ where $U_s = \left(U_s^d, U_s^c\right)$. We factor analyze them under assumptions like those invoked in matrix (21) with two or more of these elements dependent solely on θ_1, an additional two or more elements dependent solely on (θ_1, θ_2) and so forth but at least three final elements dependent on θ_K. There are a total of $\overline{A} \times \overline{R}$ outcomes in each state where $\overline{R}$ is the number of outcome measures in each state at each age (*e.g.*, wages, employment, occupation), there are $\overline{M}$ non-state-contingent measurements and $\tilde{\varepsilon}_W$ is a scalar. Thus L in (21) is $(\overline{A} \times \overline{R}) + \overline{M} + 1$ in dimension for each system $s, s = 1, ..., \overline{S}$.

We write the unobservables in factor structure form

$$\begin{aligned} U_{s,a} &= \boldsymbol{\alpha}_{s,a}' \boldsymbol{\theta} + \varepsilon_{s,a} \text{ with } s = 1, ..., \overline{S} \quad a = 1, ..., \overline{A} \\ U_m &= \boldsymbol{\alpha}_m' \boldsymbol{\theta} + \varepsilon_m \text{ with } m = 1, ..., N_m \\ \tilde{\varepsilon}_W &= \boldsymbol{\gamma}' \boldsymbol{\theta} + \varepsilon_I. \end{aligned}$$

The $\boldsymbol{\alpha}_{s,a}$ may be different across s-states so that each s system may depend on different elements of $\boldsymbol{\theta}$. The $\boldsymbol{\alpha}_m$ are not, nor is the $\boldsymbol{\gamma}$. There may be multiple measurements of outcomes so in principle $\boldsymbol{\alpha}_{s,a}$ may be a matrix and $\varepsilon_{s,a}$ a vector of mutually independent components. Our empirical analysis is for the vector case.

The choice of how to select the blocks of (21) may appear to be arbitrary, but in many applications there are natural orderings. Thus in the empirical work reported below we estimate a two factor model. We have a vector of five test scores that proxy latent ability (θ_1). The state contingent outcomes (earnings) equations and choice equations plausibly depend on both θ_1 and $\theta_{2.}$ In many applications there are often natural allocations of factors to various measurements. However, to avoid arbitrariness a carefully reasoned defense of any allocation is required. We now formalize identification in this system.

Theorem 4 *Under the normalizations on the factor loadings of the type in (21) for one system s under the conditions of Theorems 1-3, given the normalizations for the unobservables for the discrete components and given at least* $2K + 1$ *measurements* (Y, M, I)*, the unrestricted factor loadings and the variances of the factors* $(\sigma_{\theta_i}^2, i = 1, ..., K)$ *are identified for all systems.*

Proof: The proof is implicit in the discussion surrounding equation (21). ■

Observe that since the $\sigma_{\theta_i}^2, i = 1, ..., K$ are identified in one system, normalizations of specific factor loadings to unity are only required in that system since we can apply the knowledge of these variances to the other systems.[26] Thus for the other systems (values of the state other than s) we do not need to normalize any factor loading to unity.

We can also nonparametrically identify the densities of the uniquenesses and the factors. This follows from mutual independence of the θ_i, $i = 1, ..., K$ and an application of Kotlarski's Theorem (1967). We first state Kotlarski's Theorem and then we apply it to our problem.

Write $(\{U_m\}_{m=1}^{N_m}, \{U_{s,a}\}_{a=1}^{\overline{A}}, \tilde{\varepsilon}_W)$ in vector form as T^s. Order the vectors so that the first B_1 (≥ 2) elements depend only on θ_1, the next $B_2 - B_1$ (≥ 2) elements depend on (θ_1, θ_2) and so forth. Let T_1^s and T_2^s be the first two elements of T^s. (This is purely a notational convenience). We order the elements of T^s so that the first block depends solely on θ_1, (assuming that there are B_1 such measurements) the second block depends solely on θ_1, θ_2 (there are $B_2 - B_1$ such measurements) and so forth, following the convention established in equation (21). We require $B_1 \geq 2$, $B_2 - B_1 \geq 2$, and $B_K - B_{K-1} \geq 3$.

Theorem 5 *If*

$$T_1^s = \theta_1 + v_1$$

and

$$T_2^s = \theta_1 + v_2$$

and $\theta_1 \perp\!\!\!\perp v_1 \perp\!\!\!\perp v_2$, *the means of all three generating random variables are finite,* $E(v_1) = E(v_2) = 0$, *and the conditions of Fubini's theorem are satisfied for each random variable, and the random variables possess nonvanishing (*a.e.*) characteristic functions, then the densities of* $(\theta_1, v_1, v_2), g(\theta_1), g_1(v_1), g_2(v_2)$, *respectively, are identified.*

Proof: Kotlarski (1967). See also Rao (1992). ■

Applied to our context, consider the first two equations of T and suppose that the components depend only on θ_1. We use our notation for the factor loadings to write

$$\begin{aligned} T_1^s &= \lambda_{11}^s \theta_1 + \varepsilon_1^s \text{ where } \lambda_{11}^s = 1 \\ T_2^s &= \lambda_{21}^s \theta_1 + \varepsilon_2^s \text{ where } \lambda_{21}^s \neq 0. \end{aligned}$$

Here we use a notation associating the subscript of ε_i^s with its position in the T^s vector. Applying Theorem 4, we can identify λ_{21}^s (subject to the normalization $\lambda_{11}^s = 1$).[27] Thus we can rewrite these equations as

$$\begin{aligned} T_1^s &= \theta_1 + \varepsilon_1^s \\ \frac{T_2^s}{\lambda_{21}^s} &= \theta_1 + \varepsilon_2^{*,s}, \end{aligned}$$

where $\varepsilon_2^{*,s} = \varepsilon_2^s / \lambda_{21}^s$. Applying Kotlarski's Theorem, we can nonparametrically identify the densities $g(\theta_1), g_1(\varepsilon_1^s)$ and $g_2(\varepsilon_2^{*,s})$. Since we know λ_{21}^s we can identify $g(\varepsilon_2^s)$. Let B_1 denote the number of measurements (elements of T^s) which depend only on θ_1. Proceeding through the first B_1 measurements, we can identify $g(\varepsilon_i^s)$, $i = 1, ..., B_1$.

Proceeding to equations $B_1 + 1$ and $B_1 + 2$ (corresponding to the first two measurements in the next set of equations that depend on θ_1 and θ_2), we may use the normalization adopted in Theorem 4 to write

$$\begin{aligned} T_{B_1+1}^s &= \lambda_{B_1+1,1}^s \theta_1 + \theta_2 + \varepsilon_{B_1+1}^s \\ T_{B_1+2}^s &= \lambda_{B_1+2,1}^s \theta_1 + \lambda_{B_1+2,2}^s \theta_2 + \varepsilon_{B_1+2}^s. \end{aligned}$$

Rearranging, we may write these equations as

$$\begin{aligned} T^s_{B+1} - \lambda^s_{B_1+1,1}\theta_1 &= \theta_2 + \varepsilon^s_{B_1+1} \\ \frac{T^s_{B+2} - \lambda^s_{B_1+2,1}\theta_1}{\lambda^s_{B_1+2,2}} &= \theta_2 + \varepsilon^{*,s}_{B_1+2} \end{aligned}$$

where $\varepsilon^{*,s}_{B_1+2} = \frac{\varepsilon^s_{B_1+2}}{\lambda^s_{B_1+2,2}}$, and the $\varepsilon^s_{B_1+1}$ and $\varepsilon^{*,s}_{B_1+2}$ are mutually independent. Hence by Theorem 5, we can identify densities $g(\theta_2), g(\varepsilon^s_{B_1+1}), g(\varepsilon^s_{B_1+2})$. Exploiting the structure (21), we can proceed sequentially to identify the densities of θ, $g(\theta_i), i = 1, ..., K$ and the uniqueness, $g(\varepsilon^s_i)$ for all the components of vector T^s. For the components of ε^s_i corresponding to discrete measurements, we do not identify the scale. Armed with knowledge of the densities of the θ_i and the factor loadings for other values of s, we can apply standard deconvolution methods to nonparametrically identify the uniqueness of the ε_i's for the other systems. Thus we can nonparametrically identify the error terms for the model. Notice that in principle we can estimate separate distributions of the θ_i for each s system and thus can test the hypothesis of equality of these distributions across systems.

The essential idea in this paper is to obtain identification of the joint counterfactual distributions through the dependence across s of $Y_s = (Y^d_s, Y^c_s)$ on the common factors that also generate M or I. In this sense measurements and choices are both sources of identifying information, and can be traded off in terms of identification. We next apply our framework to a well-posed economic model.

6 Generalizing The Willis-Rosen Model

We revisit Willis and Rosen's application of the Roy model (1979) to the economics of education, adding uncertainty, nonpecuniary net returns to schooling and identifying counterfactual distributions of gross and net returns. In this paper the outcomes are utility outcomes, present value outcomes and rates of return.

Suppose that agents cannot lend or borrow and possess log preferences (utility = $\ln C$, where C is consumption). Suppose that agents are choosing between high school and college so $\overline{S} = 2$. The utility of attending college is

$$V(1) = \sum_{a=0}^{A} \frac{\ln Y^1_a}{(1+\rho)^a} - \ln P$$

where $\ln P$ is the "cost" of going to school. These include tuition costs and the psychic benefits from working in sector 1 (relative to sector 0). Thus costs may be negative. ρ is a subjective rate of time preference. The utility of completing only high school is

$$V(0) = \sum_{a=0}^{A} \frac{\ln Y^o_a}{(1+\rho)^a}$$

where Y^1_a and Y^0_a are earnings from high school and college, respectively, at age a. The psychic costs or benefits in logs for high school are normalized to zero. We can only identify relative psychic "costs" or benefits.

Latent variables and costs are generated by a factor structure. The equations are:

$$\begin{aligned} \ln Y_a^j &= \mu^j(X) + \left(\boldsymbol{\alpha}_a^j\right)' \boldsymbol{\theta} + \varepsilon_a^j \qquad\qquad j = 0, 1,\ a = 1, ..., A. \\ \ln P &= \mu^P(Z) + \left(\boldsymbol{\alpha}^P\right)' \boldsymbol{\theta} + \varepsilon^P. \end{aligned}$$

In addition we have measurements on test scores $M = \boldsymbol{\mu}_M(x) + \alpha_M' \boldsymbol{\theta} + \varepsilon^M$, where $\boldsymbol{\theta} \perp\!\!\!\perp \left[\left(\varepsilon_{i,a}^j \right)_{i=1}^{I} {}_{,j=0}^{\ 1} {}_{,a=0}^{\ A}, \varepsilon^P \right]$.

The agent makes decisions about schooling under uncertainty about different components of the model. $\mathcal{I}_\theta$ is the information set. The expected value V of going to college is :

$$V = E\left(V\left(1\right) - V\left(0\right) \mid \mathcal{I}_\theta\right) = E_{\mathcal{I}_\theta} \left[\begin{array}{c} \sum\limits_{a=0}^{A} \frac{\mu_a^1(X) - \mu_a^0(X) + \left(\boldsymbol{\alpha}_a^1 - \boldsymbol{\alpha}_a^0\right)' \boldsymbol{\theta} + \varepsilon_a^1 - \varepsilon_a^0}{(1+\rho)^a} \\ - \left[\mu_P(Z) + \boldsymbol{\alpha}_P' \boldsymbol{\theta} + \varepsilon^P\right] \end{array} \right].$$

If future innovations in earnings $\left(\varepsilon_a^1, \varepsilon_a^0\right), a = 0, .., A$ are not known at the time schooling decisions are made but innovations in costs are known, we may write the agent's preference function as

$$V = \left(\sum_{a=0}^{A} \frac{\mu_a^l\left(X\right) - \mu_a^o\left(X\right)}{\left(1+\rho\right)^a} - \mu_P\left(Z\right) \right) + \left[\sum_{a=0}^{A} \frac{\left(\boldsymbol{\alpha}_a^1 - \boldsymbol{\alpha}_a^0\right)'}{\left(1+\rho\right)^a} - \boldsymbol{\alpha}_P' \right] E_{\mathcal{I}_\theta}\left(\boldsymbol{\theta}\right) - \varepsilon^P.$$

As we shall see, this assumption about agent knowledge of future innovations in earnings is testable. Assume that $\sigma_P = \left(Var\left(\varepsilon^P\right)\right)^{\frac{1}{2}} < \infty$. Then

$$\frac{V}{\sigma_P} = \frac{1}{\sigma_P} \left(\sum_{a=0}^{A} \frac{\mu_a^1\left(X\right) - \mu_a^0\left(X\right)}{\left(1+\rho\right)^a} - \mu_P\left(Z\right) \right) + \left(\sum_{a=0}^{A} \frac{\left(\boldsymbol{\alpha}_a^1 - \boldsymbol{\alpha}_a^0\right)'}{\left(1+\rho\right)^a} - \boldsymbol{\alpha}_P' \right) \frac{1}{\sigma_P} E_{\mathcal{I}_\theta}\left(\boldsymbol{\theta}\right) - \frac{\varepsilon^P}{\sigma_P}$$

$D_s = 1$ if $\frac{V}{\sigma_P} > 0$; $D_s = 0$ otherwise.

Specifying alternative information sets ($\mathcal{I}_\theta$) and examining the resulting fit of the model to data, we can determine which information sets agents act on. Exact econometric specifications are presented in Section 7. We test whether agents act on components of $\boldsymbol{\theta}$ that also appear in outcome equations realized after the choices are made. The estimated dependence between schooling choices and subsequent realizations of earnings enables us to identify the components in the agent's information set at the time schooling decisions are being made. This extends the method of Flavin (1981) and Hansen, Roberds and Sargent (1991) to a discrete choice setting. If agents do not act on these components, then those components are intrinsically uncertain at the time agents make their schooling decisions unless nongeneric cancellations occur.[28] Because we can identify the joint distributions of unobservables, we can answer questions Willis and Rosen could not such as: (1) How highly correlated are latent skills (utilities) across sectoral choices? (2) How much intrinsic uncertainty do agents face? (3) How important is uncertainty for explaining schooling choices? (4) What fraction of the population regrets its *ex ante* schooling choice *ex post*? We can also separate out net psychic components of the returns to schooling (the $\ln P$) from monetary components.

Observe that as a consequence of the log specification of preferences (including the additive separability of the $\boldsymbol{\theta}$ and $\boldsymbol{\varepsilon}$), mean preserving spreads in $\boldsymbol{\varepsilon}_a^j, \boldsymbol{\theta}$ and ε^P produce no change in mean utility. The probability of selection $D_s = 1$ is also invariant to mean preserving spreads in ε_a^j but not for $\boldsymbol{\theta}$ and ε^P since their variance enters the choice probability if these components are known to the agent.

In addition, a mean preserving spread in $\ln Y$ is not the same as a mean preserving spread in Y. Mean preserving spreads in Y have an effect on utility since $E(Y) = e^{\mu}E(e^{\varepsilon})$. Define the residual from the mean as H, $H = e^{\mu}e^{\varepsilon} - e^{\mu}E(e^{\varepsilon})$ so $Var(H) = e^{2\mu}\left(E\left(e^{2\varepsilon}\right) - \left[E\left(e^{\varepsilon}\right)\right]^2\right)$. A mean preserving spread keeps the mean of Y fixed at constant $k = E(Y) = e^{\mu}E(e^{\varepsilon})$.

For a perturbation in the variance of ε that changes ε to $\Delta\varepsilon$, and defining $f(\varepsilon)$ as the density of ε, locally $0 = d\mu + \frac{\left[\int \varepsilon e^{\varepsilon} f(\varepsilon) d\varepsilon\right]}{E(e^{\varepsilon})} d\Delta$ so $d\mu = -\frac{\left[\int \varepsilon e^{\varepsilon} f(\varepsilon) d\varepsilon\right]}{E(e^{\varepsilon})} d\Delta$. Moreover, because $E(\varepsilon) = 0$ and $\varepsilon e^{\varepsilon}$ is convex increasing in ε, the derivative is positive. In a log normal example, $E(e^{\varepsilon}) = e^{\frac{\sigma^2}{2}}$, $E\left(e^{2\varepsilon}\right) = e^{2\sigma^2}, Var(H) = e^{2\mu}\left(e^{2\sigma^2} - e^{\sigma^2}\right), k = e^{\mu}e^{\frac{\sigma^2}{2}}, \ln k = \mu + \frac{\sigma^2}{2}, (-d\mu) = \frac{d\left(\sigma^2\right)}{2}$ so an increase in the variance is equivalent to a decrease in the mean utility. We consider the effects of mean preserving spreads on both mean log utility and on the probability that V is positive (college is selected). We now turn to the empirical analysis of this paper.

7 Empirical Results

We use the NLSY data for white males described in Appendix B and augmented with the PSID data to estimate the Willis-Rosen Model. Main features of the data are presented in Table 2. We focus on two schooling decisions; graduating from a four year college or graduating from high school. We thus abstract from the full multiplicity of choices of schooling. This is clearly a bold simplification but it allows us to focus on the main points of this paper.

As a measurement system (M) for cognitive ability we use five components of the *ASVAB* test battery (arithmetic reasoning, word knowledge, paragraph composition, math knowledge and coding speed). We dedicate the first factor (θ_1) to the ability measurement system and exclude the other factors from that system (recall the normalizations in equation (21)). We include family background variables as additional covariates in the *ASVAB* test equations (the $\boldsymbol{\mu}_M(X)$).

To simplify the empirical analysis, we divide the lifetimes of individuals into two periods. The first period covers ages 19 to 29, and the second covers ages 30 to 65. We compute annual earnings by multiplying the hourly wage by hours worked each year for each individual.[29] We impute missing wages and project earnings for the ages not observed in the NLSY data using the procedure described in Appendix B. The NLSY data do not contain information on the full life cycle of earnings. We project the missing NLSY earnings using estimates of lifetime earnings from the PSID data.

Tables 2a-b present the sample statistics. They show that while college graduates have higher earnings than high school graduates, all of the gain to attending college comes after age 30. College graduates also have much higher test scores and come from better family backgrounds than high school graduates. They are more likely to live in locations where a college is present and where college tuition is lower.

In the notation of Section 5, $\bar{S} = 2$ (two choices), $\bar{R} = 1$ (there is one outcome per person, earnings), $\bar{M} = 5$ (there are five test scores that are generated solely by θ_1) and $\bar{A} = 2$ (there are two periods in the life cycle). In addition, there is utility index I. The test scores depend solely on θ_1. The outcomes and index are allowed to depend on (θ_1, θ_2). Since $K = 2$, assuming non-zero factor loadings, we satisfy the conditions for identification presented in Theorem 4. We have five measurements generated solely by θ_1. There are three measurements generated by θ_1 and θ_2 for each schooling level. (Outcomes and choices are defined for each choice system). Exclusion restrictions are given in Table 2c along with specification of each of the equations. Tuition and family background identify the parameters of the schooling equations. Local labor market variables identify the parameters of utility equations. Assuming that test scores are continuous outcomes, no exclusions are needed for identification of the test score equations and their distribution.

In this section, to facilitate the exposition we denote the college state (choice 1) by c, while high school (choice 0) is denoted by h. We model log earnings (utility of earnings) at each age as:

$$\ln Y_{a,s} = \delta_{a,s} + X'\boldsymbol{\beta}_{a,s} + \eta_{1,s} * \text{experience}_a + \eta_{2,s} * \text{experience}_a^2 + \boldsymbol{\alpha}'_{a,s}\boldsymbol{\theta} + \varepsilon_{a,s} \tag{23}$$

where $Y_{a,s}$ is earnings in period (age) a if the schooling level is s, X is a vector of covariates, $\boldsymbol{\theta}$ is a vector of factors and $\eta_{1,s}$ and $\eta_{2,s}$ are calculated by the procedure described in Appendix B. We compute the present value of log earnings (lifetime utility) in the first period (ages 19 to 29) and in the second period (ages 30 to 65). Let $V_{1,s}$ be the period 1 gross utility of achieving schooling level s, and $V_{2,s}$ be the period 2 gross utility of obtaining schooling level s. Using (23), we write the gross utilities as

$$\begin{aligned}
V_{1,s} &= \bar{\delta}_{1,s} + X'\bar{\boldsymbol{\beta}}_{1,s} + \bar{\boldsymbol{\alpha}}'_{1,s}\boldsymbol{\theta} + \bar{\varepsilon}_{1,s} \\
V_{2,s} &= \bar{\delta}_{2,s} + X'\bar{\boldsymbol{\beta}}_{2,s} + \bar{\boldsymbol{\alpha}}'_{2,s}\boldsymbol{\theta} + \bar{\varepsilon}_{2,s}.
\end{aligned}$$

These are the outcome equations for the model that we estimate. To see this, notice that

$$\begin{aligned}
V_{1,s} &= \sum_{a=19}^{A_1} \frac{\ln Y_{a,s}}{(1+\rho)^a} \\
&= \sum_{a=19}^{A_1} \frac{\delta_{a,s} + X'\boldsymbol{\beta}_{a,s} + \boldsymbol{\alpha}'_{a,s}\boldsymbol{\theta} + \varepsilon_{a,s} + \eta_{1,s} * \text{experience}_a + \eta_{2,s} * \text{experience}_a^2}{(1+\rho)^a} \\
&= \sum_{a=19}^{A_1} \frac{\delta_{a,s} + X'\boldsymbol{\beta}_{a,s} + \eta_{1,s} * \text{experience}_a + \eta_{2,s} * \text{experience}_a^2}{(1+\rho)^a} + \left[\sum_{a=19}^{A_1} \frac{\boldsymbol{\alpha}'_{a,s}}{(1+\rho)^a}\right]\boldsymbol{\theta} + \sum_{a=19}^{A_1} \frac{\varepsilon_{a,s}}{(1+\rho)^a} \\
&= \bar{\delta}_{1,s} + X'\bar{\boldsymbol{\beta}}_{1,s} + \bar{\boldsymbol{\alpha}}'_{1,s}\boldsymbol{\theta} + \bar{\varepsilon}_{1,s}
\end{aligned}$$

where

$$
\begin{aligned}
A_1 &= 29 \\
\rho &= 0.03 \text{ (the prespecified discount rate)} \\
\bar{\delta}_{1,s} &= \sum_{a=19}^{A_1} \frac{\delta_{a,s} + \eta_{1,s} * \text{experience}_a + \eta_{2,s} * \text{experience}_a^2}{(1+\rho)^a} \\
\bar{\boldsymbol{\beta}}_{1,s} &= \sum_{a=19}^{A_1} \frac{\boldsymbol{\beta}_{a,s}}{(1+\rho)^a} \\
\bar{\boldsymbol{\alpha}}'_{1,s} &= \sum_{a=19}^{A_1} \frac{\boldsymbol{\alpha}'_{a,s}}{(1+\rho)^a} \\
\bar{\varepsilon}_{1,s} &= \sum_{a=19}^{A_1} \frac{\varepsilon_{a,s}}{(1+\rho)^a}
\end{aligned}
$$

and terms for the second period of data (30-65) are defined analogously. The "cost" or psychic net return of going to college is written as:

$$\ln P = \delta_P + Z'\gamma + \boldsymbol{\alpha}'_P\boldsymbol{\theta} + \varepsilon_P.$$

These "costs" can be negative as they entail both psychic and tuition components. Assuming that the agents know X, Z, $\boldsymbol{\theta}$ and ε_P, the criterion for the choice of schooling is:

$$
\begin{aligned}
V &= E\left(V_{1,c} + V_{2,c} - V_{1,h} - V_{2,h} | X, \boldsymbol{\theta}\right) - E(\ln P | Z, X, \boldsymbol{\theta}, \varepsilon_P) \\
&= \bar{\delta}_{1,c} + X'\bar{\boldsymbol{\beta}}_{1,c} + \bar{\boldsymbol{\alpha}}'_{1,c}\boldsymbol{\theta} + \bar{\delta}_{2,c} + X'\bar{\boldsymbol{\beta}}_{2,c} + \bar{\boldsymbol{\alpha}}'_{2,c}\boldsymbol{\theta} - \bar{\delta}_{1,h} - X'\bar{\boldsymbol{\beta}}_{1,h} - \bar{\boldsymbol{\alpha}}'_{1,h}\boldsymbol{\theta} - \bar{\delta}_{2,h} - X'\bar{\boldsymbol{\beta}}_{2,h} - \bar{\boldsymbol{\alpha}}'_{2,h}\boldsymbol{\theta} \\
&\quad -\delta_P - Z'\boldsymbol{\gamma} - \boldsymbol{\alpha}'_P\boldsymbol{\theta} - \varepsilon_P \\
&= \left(\bar{\delta}_{1,c} + \bar{\delta}_{2,c} - \bar{\delta}_{1,h} - \bar{\delta}_{2,h} - \delta_P\right) + X'\left(\bar{\boldsymbol{\beta}}_{1,c} + \bar{\boldsymbol{\beta}}_{2,c} - \bar{\boldsymbol{\beta}}_{1,h} - \bar{\boldsymbol{\beta}}_{2,h}\right) - Z'\boldsymbol{\gamma} \\
&\quad + \left(\bar{\boldsymbol{\alpha}}'_{1,c} + \bar{\boldsymbol{\alpha}}'_{2,c} - \bar{\boldsymbol{\alpha}}'_{1,h} - \bar{\boldsymbol{\alpha}}'_{2,h} - \boldsymbol{\alpha}'_P\right)\boldsymbol{\theta} - \varepsilon_P.
\end{aligned}
$$

Individuals go to college if $V > 0$. We test (and do not reject) the hypothesis that at the time they make their college decision agents know their cost function and both factors $\boldsymbol{\theta}$, but not the uniquenesses in the outcome equations. These expressions can be modified in an obvious way to accommodate other information sets.

The test score equations have a similar structure. Let T_j be test score j:

$$T_j = X'\boldsymbol{\omega}_j + \boldsymbol{\alpha}'_{\text{test}_j}\boldsymbol{\theta} + \varepsilon_{\text{test}_j}$$

where X is the vector of covariates in the test score equation, and $\boldsymbol{\omega}_j$ is the covariate vector. The distributions of the $\boldsymbol{\theta}$ and $\boldsymbol{\varepsilon}$ are nonparametrically identified under the assumptions supporting Theorems 1-5. In this paper, we assume that each factor is generated by a mixture of normals distribution,

$$\theta_k \sim \sum_{j=1}^{J_k} p_{k,j}\phi\left(f_k|\mu_{j,k},\tau_{j,k}\right), \qquad k=1,\ldots,K. \tag{24}$$

Mixtures of normals with a large enough number of components approximate any distribution of $\boldsymbol{\theta}_k$ and the $\boldsymbol{\varepsilon}$ arbitrarily well (Ferguson, 1983). We assume that the ε's are normal although in principle they are nonparametrically identified from the analysis of Theorem 5.

We estimate the model using Markov Chain Monte Carlo methods as described in Appendix C for 55,000 iterations, discarding the first 5,000 iterations to allow the chain to converge to its stationary distribution. We retain every 10th of the remaining 50,000 iterations for a total of 5,000 iterations.[30] The Markov Chain mixes well with most autocorrelations dying out at around lag 25 to 50.

We estimate models with one factor and with two factors. The estimated coefficients are presented as Tables A1 through A5 in the supplementary tables on the website (http://lily.src.uchicago.edu/CHH_estimating.html). The two factor model specifies that the first factor only appears in test scores and choice equations while the second factor appears in all equations. No additional factors are necessary to fit our data. Thus we conclude that the innovations in the earnings process $\left(\varepsilon_a^j\right)$ are not in the agent's information set at the time schooling decisions are made. If they were, they would be an additional source of covariance (*i.e.*, they would generate additional factors) between the choice equation and future earnings. If we use only one factor that enters in all equations, the quality of the fit is much poorer (results available on request). From this testing procedure we infer that agents know both components of $\boldsymbol{\theta}$ at the time they enroll in college. Figure 1 shows the fit of the density of the present value of log earnings (or lifetime utility of earnings excluding psychic costs and benefits) for everyone in the population. It graphs the actual and predicted densities of gross utility. The fit is very good. Results for each schooling group are available in the supplement on the website and are equally good (χ^2 goodness of fit tests are passed overall as well as for the distribution of utility for each schooling group; see Table A6). In order to achieve this good fit it is necessary to allow for non-normal factors. Figure 2 shows the density of each of the estimated factors and compares them with a benchmark normal with the same mean and standard deviation. Neither factor is normal.[31] There is evidence of selection on ability (factor 1), with the less able less likely to attend college. There is weaker evidence of selection on factor 2 (see graphs A-1 and A-2 posted at the website).

Tables 3a-b presents the factor loadings in the outcome, choice and measurement equations.[32] Both factors have a positive effect on gross utility for both schooling levels in each period and on schooling attainment (the I). Factor 1 explains most of the variance in the test score system (see Table 3b) while factor 2 explains most of the variance in the utility outcome system (see Table 3a). The return to college in terms of gross utility (gross utility differences) is given by:

$$\begin{aligned} V_{1,c}+V_{2,c}-V_{1,h}-V_{2,h} \quad &= \quad \left(\bar{\delta}_{1,c}+\bar{\delta}_{2,c}-\bar{\delta}_{1,h}-\bar{\delta}_{2,h}\right)+X'\left(\bar{\boldsymbol{\beta}}_{1,c}+\bar{\boldsymbol{\beta}}_{2,c}-\bar{\boldsymbol{\beta}}_{1,h}-\bar{\boldsymbol{\beta}}_{2,h}\right) \\ &\quad + \left(\bar{\boldsymbol{\alpha}}'_{1,c}+\bar{\boldsymbol{\alpha}}'_{2,c}-\bar{\boldsymbol{\alpha}}'_{1,h}-\bar{\boldsymbol{\alpha}}'_{2,h}\right)\boldsymbol{\theta}+\left(\bar{\varepsilon}_{1,c}+\bar{\varepsilon}_{2,c}-\bar{\varepsilon}_{1,h}-\bar{\varepsilon}_{2,h}\right). \end{aligned}$$

Both factors raise returns (see the base of Table 3a). While the second factor explains much more of the variance in utility than the first factor, the first factor explains more of the variance in returns than the second factor although it only explains 30% of the variance in returns. We infer that agents know $\boldsymbol{\theta}$ (the factors) based on the superior fit of a model that includes nonzero factor loadings on both factors in the choice equation but not the innovations in outcomes (the $\boldsymbol{\varepsilon}$'s in the outcome equations) at the time they make their schooling decisions.

Our results indicate that the unpredictability in gross utility *gains* (*i.e.* differences) of going to college is much larger than the unpredictability in utility levels. Both factors have a negative impact on "costs" (the factor loadings are positive in the "cost" or psychic return function). Therefore, both factors positively influence the likelihood of going to college since both contribute positively to returns and negatively to costs.

Figure 3 plots the estimated factual and counterfactual gross college utility densities for college graduates and high school graduates, respectively (see Figure A3 on the website for the corresponding figure for high school utility). College graduates have the highest level of gross utility both as high school graduates and as college graduates. They also have the highest gross gains of going to college as demonstrated in Figure 4.[33,34] Figure 5 presents the marginal treatment effect as defined in equation (6) using utils as the outcome. This is the gross gain in utils of going to college as a function of ε_W, which is an index of variables that increase the likelihood of enrollment in college. It shows that individuals who are likely to enroll in college have higher returns to college than those who are unlikely to enroll in college who have lower values of ε_W. Figure 5 also shows the distribution of ε_W in the population. Most of the mass of this distribution is at values of ε_W around 0. Many individuals have negative gross utility returns (excluding psychic benefits of going to college). Even among those deciding to go to college, 39.53% would have higher utility (ignoring psychic components) had they not gone to college. There is a definite falloff in utility gains as college enrollment is expanded to the less college prone. Table 4 confirms Figure 3 and shows that college graduates have higher potential high school and college utility than high school graduates in high school and in college (these are gross utilities). Table 5 shows that the gross returns of going to college are higher for those who choose to go to college. These results are expected given the pattern shown in Figure 4. The returns for attending college for the average high school graduate are negative. The returns to college for the individual at the margin ($V = 0$) are about 0.59% of total high school utility. Since these individuals are exactly at the margin, these gains correspond exactly to the cost they are facing. Once we account for the nonmonetary costs and benefits of going to college (net returns reported in the bottom two rows of Table 5) the relative returns of going to college become more negative for high school graduates and more positive for college graduates. Since $\ln P$ can be allocated as either a cost or a return, there are two ways to compute returns depending on whether $\ln P$ is treated as a cost (row 2) or a return (row 3). We present two sets of net return estimates depending on how "costs" or "gains" ($\ln P$) are allocated. These are bounds since the actual allocation between cost and benefit is indeterminate.

The patterns of Figures 3-5 are essentially reproduced for present value of earnings in Figures 6-8. Table 6 shows that college graduates have earnings 57.6% higher than they would have had (or $608,372 higher, on average) if they did not go to college. High school graduates have a gross gain of 43% (or $362,987) if they go to college. Notice that even though the utility gains of going to college are negative for high school graduates, the money returns are positive and large. Table 7 shows that even though 39.66% of the persons going to college would have had a higher utility in high school than in college (ignoring psychic gains), only 6.9% of this population had higher earnings in high school than in college. Once we account for psychic benefits, the proportion of college students regretting their decisions is roughly the same whether we measure regret in present value or utils. This shows the importance of accounting for psychic returns in analyzing schooling choices. Among high school graduates, 95.90% do not regret not going to college (measured in utils), but 85.26% regret the decision financially. The marginal treatment effect has the same general shape when present values of earnings are used instead of gross utility (see Figure 8).

Table 8 shows the probability of being in decile i of the college potential discounted earnings distribution conditional on being in decile j of the high school potential earnings distribution. (These are gross earnings.) It shows that neither an independence assumption across counterfactual outcomes, which is the Veil of Ignorance assumption used in applied welfare theory, (see, *e.g.*, Sen, 1973) or in aggregate income inequality decompositions (DiNardo, Fortin, and Lemieux, 1996) nor a perfect ranking assumption, which are sometimes used to construct counterfactual joint distributions of outcomes, (see e.g. Heckman, Smith, and Clements, 1997 or Athey and Imbens, 2002) are satisfied in the data. There is a strong positive dependence between potential outcomes in each counterfactual state, but there is not perfect dependence. There are substantial nonzero elements outside the diagonal. We get similar results for utils (discounted log earnings). See Table A-13 at our website.

We have already shown that there is a large dispersion in the distribution of utilities, utility returns, earnings, and earnings returns to college. However, this dispersion can be due to heterogeneity that is known at the time the agent makes schooling decisions, or it can be due to heterogeneity that is not predictable by the agent at that time. Figure 9 plots the densities of the unforecastable component of college gross utilities at the time college decisions are made for fixed X values, under three different information sets. (The X are fixed at their means.) The solid line corresponds to the case where the agent does not know his factor ($\boldsymbol{\theta}$) nor his innovations (the $\boldsymbol{\varepsilon}$'s in the outcome equations). The other two lines correspond respectively to the cases where the agent knows θ_2 only, or both θ_1 and θ_2.[35] Knowledge of θ_2 dramatically decreases the uncertainty faced, but knowledge of factor 1 (associated with cognitive ability) has only a small effect on the amount of uncertainty faced by the agent. We obtain a similar figure in terms of gross utility in high school.[36] However, even though knowledge of θ_2 reduces dramatically the amount of uncertainty faced in terms of levels of gross utility in each counterfactual state, it has only a small effect on the uncertainty faced in terms of *returns* (see Figure 10). Table 9 reports the variances of gross and net utility and gross and net present value of earnings under different information sets of agents. Giving agents more information (knowledge of factors) reduces the variance in utilities or present values as perceived by agents. However, reducing uncertainty barely budges the forecast returns to schooling measured in dollars or utils–the message of Figure 10. Analogous results are obtained for present value of earnings. See Figures A-15 and A-16 posted at our website.

The fact that a two factor model is adequate to fit the data implies that the agents cannot forecast future shocks of log earnings $(\bar{\varepsilon}_{1,c}, \bar{\varepsilon}_{2,c}, \bar{\varepsilon}_{1,h}, \bar{\varepsilon}_{2,h})$ at the time they make their schooling decision. (If they did, they would enter as additional factors in the estimated model.) Even though the factors ($\boldsymbol{\theta}$) explain most of the variance in levels of utilities, they explain less than half of the variance in returns, which may lead the reader to conclude that the reason so many college graduates would have higher gross utility in high school than in college (39%) is because they cannot accurately forecast their returns of going to college. However this is not the case. As shown in Table 7 once we account for psychic benefits or costs of attending college (P) relative to attending high school, only 8% of college graduates regret going to college. This suggests a substantial part of the gain to college is due to non-pecuniary components. Furthermore, Table 10 shows that if individuals had knowledge of $(\bar{\varepsilon}_{1,c}, \bar{\varepsilon}_{2,c}, \bar{\varepsilon}_{1,h}, \bar{\varepsilon}_{2,h})$, keeping their average expected earnings the same, very few of them would change their schooling decision. Uncertainty in gains to schooling is substantial but knowledge of this uncertainty has a very small effect on the choice of schooling because the variance of gains is so much smaller than the variance of psychic costs or benefits, and it is the latter that drives most of the heterogeneity in schooling decisions. In addition, there is uncertainty about the level of both college and high school earnings. See the variances reported for each in Table 9. The uncertainty in the return comes from both sources although the literature emphasizes the uncertainty in college earnings. When conducting this experiment, we make sure that the average expected earnings are the same because a mean preserving reduction in the uncertainty faced by the agents in terms of utility is not the same as a mean preserving change in uncertainty in terms of levels of earnings (see Appendix D).[37] In particular a change in the variance of $(\bar{\varepsilon}_{1,c}, \bar{\varepsilon}_{2,c}, \bar{\varepsilon}_{1,h}, \bar{\varepsilon}_{2,h})$ would not change the expected utility in each schooling level but would change expected earnings in each schooling level. The numbers reported in Table 10 take this into account. When agents know their $(\bar{\varepsilon}_{1,c}, \bar{\varepsilon}_{2,c}, \bar{\varepsilon}_{1,h}, \bar{\varepsilon}_{2,h})$, they face less uncertainty. Knowing these components is equivalent to setting $\Delta = 0$ in the expression at the end of Section 6, a special case of mean preserving shrinkage where variances are set to zero. The expected utility at each schooling level increases.[38]

8 Some Evidence on an Educational Reform

Using the estimated model, we evaluate the effect of a full subsidy to college tuition. We move beyond the Veil of Ignorance which is based on an anonymity assumption and evaluates reforms considering only their overall impact on inequality, to consider which individuals are benefited by the reform. We consider only partial equilibrium treatment effects and do not consider the full cost of financing the reforms. Table 4 shows the average lifetime gross utility of participants before the policy change and Table 5 shows their pre-policy average return to college. These tables compare these levels and returns with what the marginal participant attracted into schooling by the policy would earn. The marginal person has lower utility in college and lower returns to college than the average person in college (also see Figure 5). Since the policy affects the schooling decisions of the individuals at the margin, the policy will produce a decline in the quality of college graduates after the policy is implemented, since the new entrants are of lower average quality than the incumbents.

Despite the substantial size of the policy changes we consider, the induced effects on participation are small. The full tuition subsidy only increases graduation from four-year college by 4%.[39] The policies operate unevenly over the deciles of the initial outcome distribution. Figure 11 shows the proportion of high school people in each decile of the high school present value of earnings distribution induced to graduate from four-year college by the tuition subsidy. The figure shows that providing a free college education mostly affects people at the top end of the high school earnings distribution.[40] The policy does not benefit the poor. A calculation based on the Veil of Ignorance using the Gini coefficient would show no effect of the policy up to two decimal points. Our analysis relaxes the Veil of Ignorance, and lets us study the impact of policies on persons at different positions of the income distribution. It goes beyond the counterfactual simulations used in the inequality literature (see, e.g. DiNardo, Fortin and Lemieux, 1996) to account for self selection by agents into sectors in response to policy changes.

9 Summary and Conclusions

This paper uses low dimensional factor models to generate counterfactual distributions of potential outcomes. It extends matching by allowing some of the variables that determine the conditional independence assumed in matching to be unobserved by the analyst. Semiparametric identification is established.

We apply our methods to a problem in the economics of education. We extend the Willis-Rosen model to explicitly account for dependence in potential outcomes across potential schooling states, to account for psychic benefits in the return to schooling and to measure the effect of uncertainty on schooling choices. We extend the framework of Flavin (1981) and Hansen, Roberds and Sargent (1991), who estimate the impact of uncertainty on consumption choices to a discrete choice setting to estimate agent information sets. Our framework extends the inequality decomposition analysis of DiNardo, Fortin, and Lemieux (1996) to account for self selection in the choice of sectors.

Our analysis reveals substantial heterogeneity in the returns to schooling, much of which is unpredictable at the time schooling decisions are made. We also find a substantial non-pecuniary return to college. Although there is substantial uncertainty in forecasting returns at the time schooling decisions are made, eliminating it has modest effects on schooling choices. Uncertainty is inherent in both college and high school outcomes at the time schooling decisions are made. In addition, nonpecuniary factors play a dominant role in schooling choices. The assumption of perfect ranking of potential outcome across alternative choices is soundly rejected, although potential outcomes are strongly positively correlated.

We simulate a tuition reduction policy to determine who benefits and loses from it. We go beyond the Veil of Ignorance to see which persons are affected by the policy. The policy favors those at the top of the income distribution. This simulation illustrates the power of our method to lift the Veil of Ignorance, and to count the losers and gainers from any policy initiative.

Appendix A : Proofs of Theorems

Proof of Theorem 1: The case where M consists of purely continuous components is trivial. We observe M^c for each X and can recover the marginal distribution for each component. Recall that M is not state dependent.

For the purely discrete case, we encounter the usual problem that there is no direct observable counterpart for $\mu_m^d(X)$. Under (A-1)-(A-5), we can use the analysis of Manski (1988) to identify the slope coefficients $\beta_{l,m}^d$ up to scale, and the marginal distribution of $U_{l,m}^d$. From the assumption that the mean (or median) of $U_{l,m}^d$ is zero, we can identify the intercept in $\beta_{l,m}^d$. We can repeat this for all discrete components. Thus coordinate by coordinate we can identify the marginal distributions of $U_m^c, \widetilde{U}_m^d, \mu_m^c(X)$ and $\widetilde{\mu}_m^d(X)$, the latter up to scale ("$\sim$" means identified up to scale).

To recover the joint distribution write:

$$\Pr\left(M_c \leq m_c, M_d = (0,...,0) \mid X\right) = F_{U_m^c, \widetilde{U}_m^d}\left(m_c - \mu_m^c(X), -\widetilde{\mu}_m^d(X)\right)$$

by assumption (A-2). To identify $F_{U_m^c, \widetilde{U}_m^d}(t_1, t_2)$ for any given evaluation points in the support of $(U_m^c, \widetilde{U}_m^d)$, we know the function $\widetilde{\mu}_m^d(X)$ and using (A-3) we can find an X where $\widetilde{\mu}_m^d(X) = t_2$. Let $\widehat{x}$ denote this value, so $\widetilde{\mu}_m^d(\widehat{x}) = t_2$. In this proof, t_1, t_2 may be vectors. Thus

$$\Pr\left(M_c \leq m_c, M_d = (0,...,0) \mid X = \widehat{x}\right) = F_{U_m^c, \widetilde{U}_m^d}\left(m_c - \mu_m^c(\widehat{x}), t_2\right)$$

Let $\widehat{m}_c = t_1 - \mu_m^c(\widehat{x})$ to obtain

$$\Pr\left(M_c \leq \widehat{m}_c, M_d = (0,...,0) \mid X = \widehat{x}\right) = F_{U_m^c, \widetilde{U}_m^d}(t_1, t_2)$$

We know the left hand side and thus identify $F_{U_m^c, \widetilde{U}_m^d}$ at the evaluation point t_1, t_2. Since (t_1, t_2) is any arbitrary evaluation point in the support of $U_m^c, \widetilde{U}_m^d$ we can thus identify the full joint distribution.■[41]

Proof of Theorem 2:

$$\Pr\left(D_1 = 1 \mid Z, Q_1\right) = \Pr\left(\frac{c_1(Q_1) - \varphi(Z)}{\sigma_W} > \frac{\varepsilon_W}{\sigma_W}\right)$$

Under (A-1), (A-2), (A-6), (A-7) and (A-9), it follows that $\frac{c_1(Q_1)-\varphi(Z)}{\sigma_W}$ and $F_{\widetilde{\varepsilon}_W}$ (where $\widetilde{\varepsilon}_W = \frac{\varepsilon_W}{\sigma_W}$) are identified (see Manski, 1988 or Matzkin 1992, 1993). Under rank condition (A-7), identification of $\frac{c_1(Q_1)-\varphi(Z)}{\sigma_W}$ implies identification of $\frac{c_1(Q_1)}{\sigma_W}$ and $\frac{\varphi(Z)}{\sigma_W}$ separately. Write

$$\Pr\left(D_2 = 1 \mid Z, Q_1, Q_2\right) = F_{\widetilde{\varepsilon}_W}\left(\frac{c_2(Q_2) - \varphi(Z)}{\sigma_W}\right) - F_{\widetilde{\varepsilon}_W}\left(\frac{c_1(Q_1) - \varphi(Z)}{\sigma_W}\right).$$

From the absolute continuity of $\widetilde{\varepsilon}_W$ and the assumption that the distribution function of $\widetilde{\varepsilon}_W$ is strictly increasing, we can write

$$\frac{c_2(Q_2)}{\sigma_W} = F_{\widetilde{\varepsilon}_W}^{-1}\left[\Pr\left(D_2 = 1 \mid Z, Q_1, Q_2\right) + F_{\widetilde{\varepsilon}_W}\left(\frac{c_1(Q_1) - \varphi(Z)}{\sigma_W}\right)\right] + \frac{\varphi(Z)}{\sigma_W}.$$

Thus we can identify $\frac{c_2(Q_2)}{\sigma_W}$ over its support and, proceeding sequentially, we can identify $\frac{c_s(Q_s)}{\sigma_W}$, $s = 3, .., \overline{S}$. Under (A-8) we can identify $\eta_s, s = 2, .., \overline{S}$.■ Observe that we could use the final choice ($\Pr(s = \overline{S})$) rather than the initial choice to start off the proof of identification using an obvious change in the assumptions.

Proof of Theorem 3: From (A-2), the unobservables are jointly independent of (X, Z, Q). For fixed values of (Z, Q_s, Q_{s-1}), we may vary the points of evaluation for the continuous coordinates (y_s^c) and pick alternative values of $X = \widehat{x}$ to trace out the vector $\boldsymbol{\mu}^c(X)$ up to intercept terms. Thus we can identify $\boldsymbol{\mu}_{s,l}^c(X)$ up to a constant for all $l = 1, ..., N_{c,s}$.(Heckman and Honoré, 1990). Under (A-2), we recover the same functions for whatever values of Z, Q_s, Q_{s-1} are prespecified as long as $c_s(Q_s) > c_{s-1}(Q_{s-1})$, so that there is interval of ε_W bounded above and below with positive probability. This identification result does not require any passage to a limit argument.

For values of (Z, Q_s, Q_{s-1}) such that

$$\lim_{\substack{Q_s \to \bar{Q}_s(Z) \\ Q_{s-1} \to \underline{Q}_{s-1}(Z)}} \Pr\left(D_s = 1 | Z, Q_s, Q_{s-1}\right) = 1.$$

where $\bar{Q}_s(Z)$ is an upper limit and $\underline{Q}_{s-1}(Z)$ is a lower limit, and we allow the limits to depend on Z, we essentially integrate out $\widetilde{\varepsilon}_W$ and obtain

$$\Pr(M^c \leq m^c, \widetilde{\boldsymbol{\mu}}_m^d \leq -U_m^d, U_s^c \leq y_s^c - \boldsymbol{\mu}^c(X), \widetilde{U}_s^d \leq -\widetilde{\boldsymbol{\mu}}_s^d(X))$$

We know that this probability can be achieved by virtue of the support condition of assumption (A-10).

Then proceeding as in the proof of Theorem 1, we can identify $\widetilde{\boldsymbol{\mu}}_s^d(X)$ coordinate by coordinate and we obtain the constants in $\boldsymbol{\mu}_{s,l}^c(X)$, $l = 1, ..., N_{c,s}$ as well as the constants in $\widetilde{\boldsymbol{\mu}}^d(X)$. From the assumption of mean or median zero of the unobservables. In this exercise, we use the full rank condition on X which is part of assumption (A-11).

With these functions in hand, under the full conditions of assumption (A-10) we can fix $y_s^c, y_m^c, \widetilde{\mu}_s^d, \widetilde{\mu}_m^d$, $\frac{c_s(Q_s)-\varphi(Z)}{\sigma_W}$, $\frac{c_{s-1}(Q_{s-1})-\varphi(Z)}{\sigma_W}$ at different values to trace out the joint distribution $F(U_m^c, \tilde{U}_m^d, U_s^c, \tilde{U}_s^d, \widetilde{\varepsilon}_W)$.■[42]

Appendix B: Description of the Data

We use white males from NLSY79. In the original sample there are 2439 individuals. We consider the information on these individuals from age 19 to age 35. We discard 663 individuals because they have observations missing for at least one of the covariate variables we use in the analysis. Tables 2a-b contain a description of the number of missing observations per variable. For example, we discard 50 individuals because we do not observe whether they were living in the South when they were 14 years old or not. Then we discard another 6 for not having information on whether they lived in urban area at age 14, other 5 for not reporting the number of siblings, 221 for not indicating parental education and so on, as described in Table 2a. We then restrict the NLSY sample to white males with a high school or college degree. We define high school graduates as individuals having a high school degree or having completed 12 grades and never reporting college attendance. We define participation in college as having a college degree or having completed more than 16 years in school. We exclude the oversample of poor whites. Experience is Mincer experience (age-12 if high-school graduate, age-16 for college graduate). The variables that we include in the outcome and choice equations are number of siblings, parental years of schooling, AFQT, year of birth dummies, average tuition of the colleges in the county the individual lives in at 17 (we simulate the policy change by decreasing this variable by $1000 for each individual), distance to the nearest college at 17, average local blue collar wage in state of residence at 17 (or in 1979, for individuals entering the sample at ages older than 17) and local unemployment rate in county of residence in 1979. For the construction of the tuition variable see Cameron and Heckman (2001). Distance to college is constructed by matching college location data in HEGIS (Higher Education General Information Survey) with county of residence in NLSY. State average blue collar wages are constructed using data from the BLS. For a description of the NLSY sample see BLS (2001).

In 1980, NLSY respondents were administered a battery of ten achievement tests referred to as the Armed Forces Vocational Aptitude Battery (*ASVAB*) (See Cawley, Conneely, Heckman and Vytlacil (1997) for a complete description). The math and verbal components of the *ASVAB* can be aggregated into the Armed Forces Qualification Test (AFQT) scores.[43] Many studies have used the overall AFQT score as a control variable, arguing that this is a measure of scholastic ability. We argue that AFQT is an imperfect proxy for scholastic ability and use the factor structure to capture this. We also avoid a potential aggregation bias by using each of the components of the *ASVAB* as a separate measure.

For our analysis, we use the random sample of the NLSY and restrict the sample to 1161 white males for whom we have information on schooling, several parental background variables, test scores and behavior. Distance to nearest college at each date is constructed in the following way: Take the county of residence of each individual and all other counties within the same state. The distance between two counties is defined as the distance between the center of each county. If there exists a college (2 year or 4 year) in the county of residence where a person lives then the distance to the nearest college (2 year or 4 year) variable takes the value of zero. Otherwise we compute distance (in miles) to the nearest county with a college. Then we construct distance to nearest college at 17 by using the county of residence at 17. However for people who were older than 17 in 1979 we use the county of residence in 1979 for the construction of this variable.

Tuition at age 17 is average tuition in colleges in the county of residence at 17. If there is no college in the county then average tuition in the state is taken instead. For details on the construction of this variable see Cameron and Heckman (2001).

Local labor market variables for the county of residence are computed using information in the 5% sample of the 1980 Census. For each county group in the census we compute the local unemployment rate and average wage for high school dropouts, high school graduates, individuals with some college and four year college graduates. We do not have this variable for years other than 1980 so, for each county, we assume that it is a good proxy for local labor market conditions in all the other years where NLSY respondents are assumed to be making the schooling decisions we consider in this paper.

We also use the variable log annual labor earnings. We extract this variable from the NLSY79 reported annual earnings from wages and salary. Earnings (in thousands of dollars) are discounted to 1993 using the Consumer Price Index reported by the Bureau of Labor Statistics. Missing values for this variable may occur here for two reasons: First, because respondents do not report earnings for wages/salary, and second, because the NLSY becomes biannual after 1994 and this prevents us from observing respondents when they reach certain ages. For example, because the NLSY79 was not conducted in 1995, we do not observe individuals born in 1964 when they are 31 year-old. In this case we input missing values.

To predict missing log earnings between ages 19 and 35 and extrapolate from age 36 to age 65 we pool NLSY and PSID data. From the latter, we use the sample of white males that are household heads and that are either high-school or college graduates according to the definition given above. This produces a sample of 3,043 individuals from PSID. To get annual earnings, we multiply the reported CPI-adjusted (1993 =100) hourly wage rate by the annual hours worked and divide the outcome by 1000. Then we take logs to have an NLSY-comparable variable. Similarly to NLSY, we generate the Mincerian Experience according to the rule given above. We also generate dummy variables for cohorts. The first (omitted) cohort consists of individuals born between 1896 and 1905, the second consists of individuals born between 1906 and 1915, and so on up to the last cohort which is made up of PSID respondents born between 1976 and 1985. We pool NLSY and PSID by merging the NLSY respondents in the PSID cohort born between 1956 and 1965.

Let Y_{ia} denote log earnings of agent i at age a. For each schooling choice s, we model the earnings-experience profile as

$$Y_{ia}(s) = \alpha + \beta_0 X_{ia} + \beta_1 X_{ia}^2 + D\gamma + \varepsilon_{ia} \tag{25}$$

$$\varepsilon_{ia} = \eta_i + v_{ia} \tag{26}$$

$$v_{ia} = \rho v_{ia-1} + \kappa_{ia} \tag{27}$$

where X is Mincer Experience, D is a set of dummy variables that indicate cohort, η_i is the individual effect, and κ_{ia} is white noise. In Table A-14 posted at http://lily.src.uchicago.edu/CHH_estimating.html we report the OLS estimates for $\alpha, \beta_0, \beta_1, \gamma, \rho$ based on the pooled data set.

Now, let $\hat{\varepsilon}_{ia}$ be the estimated residual of the earnings-experience profile. An estimator of the individual effect η_i is

$$\begin{aligned} \hat{\eta}_i &= \frac{1}{\sum\limits_{a=19}^{65} \phi_{ia}} \sum_{a=19}^{65} \phi_{ia}\hat{\varepsilon}_{ia}, \\ where\ \phi_{ia} &= \mathbf{1}(\text{if individual } i \text{ is observed at age } a) \end{aligned}$$

Then, we can obtain an estimator of v_{ia} by computing

$$\hat{v}_{ia} = \hat{\varepsilon}_{ia} - \hat{\eta}_i$$

Now, given $\widehat{v}_{ia}$ we can run equation (27) and then compute ρ. From this we obtain an estimator of κ_{ia} according to

$$\hat{\kappa}_{ia} = \hat{v}_{ia} - \hat{\rho}\hat{v}_{ia-1}$$

We can then predict earnings for missing observations for ages 19 to 35 and perform the extrapolation from 36 to 65 by computing for each individual

$$\begin{aligned} \hat{Y}_{ia}(s) &= \hat{\alpha} + \hat{\beta}_0 X_{ia} + \hat{\beta}_1 X_{ia}^2 + D\hat{\gamma} + \hat{\varepsilon}_{ia} \\ &= \hat{\alpha} + \hat{\beta}_0 X_{ia} + \hat{\beta}_1 X_{ia}^2 + D\hat{\gamma} + \hat{\eta}_i + \hat{\rho}\hat{v}_{ia-1} + \hat{\kappa}_{ia} \end{aligned}$$

Note that to get $\hat{\varepsilon}_{ia}$ we do not set $\hat{\kappa}_{ia}$ equal to zero. Instead, we sample ten draws from its distribution and average them for each individual, for each time period.

The next step is to get the present value of log earnings at age 19 for each agent. In order to do it we discount log earnings at each period using a discount rate of 3%. For identification purposes we then break each individual's working-life in two periods. The first one goes from age 19 to age 29. The second period goes from age 30 all the way to age 65. This produces a panel in which the first observation for each agent is the present value of log earnings from age 19 to 29 and the second is the present value of log earnings from 30 to 65. This means that lifetime present value of log earnings is just the sum of these two components. Table 2b contains descriptive statistics for the present value of log earnings for the entire working-life period and also for the two subperiods used in the analysis.

Appendix C: Markov Chain Monte Carlo Simulation Methods

Due to the complex nature of the likelihood function we will rely on Markov Chain Monte Carlo techniques to estimate the model. These are computer-intensive algorithms based on designing an ergodic discrete time continuous state Markov chain with a transition kernel having invariant measure equal to the posterior distribution of the parameter vector $\boldsymbol{\psi}$, see Robert and Casella (1999) for details. In particular, we will be using the Gibbs sampling algorithm.[44]

We first describe how the Gibbs sampler can be used to estimate models in the general set-up laid out in section 4. Let $\boldsymbol{\psi}_{s,a}$ be parameters specific to the distribution of outcomes with schooling level s at age a, let $\boldsymbol{\psi}_m$ be parameters specific to the distribution of measurements, let $\boldsymbol{\psi}_c$ be parameters specific to the distribution of schooling choice and let $\boldsymbol{\psi}_\theta$ be parameters specific to the factor distributions. Let n be the number of observations. Let the outcome matrix over all ages with schooling level s be $Y_{s,i} = (Y^c_{s,i}, Y^{*d}_{s,i})$ and the vector of measurements is M.

The complete data likelihood for completed schooling level $S = s$ is

$$f\big(M, Y_s, I, \boldsymbol{\theta}|\boldsymbol{\psi}\big) = \prod_{i:D_{i,s}=1} f\big(M_i, Y_{s,i}, I_i, \boldsymbol{\theta}_i|\boldsymbol{\psi}\big)$$

where $\boldsymbol{\psi} = \big[\boldsymbol{\psi}_{s,a}, \boldsymbol{\psi}_m, \boldsymbol{\psi}_c, \boldsymbol{\psi}_\theta\big]$, "$i$" denotes a subscript for individual i and

$$f(M_i, Y_{s,i}, I_i, \boldsymbol{\theta}_i|\boldsymbol{\psi}) = f(M_i|\boldsymbol{\psi}_m, \theta_i) \times \prod_{a=1}^{\bar{A}} f(Y_{s,a,i}|\boldsymbol{\psi}_{s,a}, \boldsymbol{\theta}_i) f(I_i|\boldsymbol{\theta}_i, \boldsymbol{\psi}_c) f(\theta_i|\boldsymbol{\psi}).$$

The complete data posterior is

$$f(M, Y,\ I, \boldsymbol{\theta}, \boldsymbol{\psi}|\text{data}) \propto \prod_{s=1}^{\bar{S}} f(\boldsymbol{\theta}, M, Y^*_s, I|\boldsymbol{\psi}) f(\boldsymbol{\psi})$$

where $Y = (Y_1, ..., Y_{\bar{S}})$.

In what follows the conditional posteriors that constitute the transition kernel of the Gibbs sampler will be derived.

Choice equations

Conditional on the factors we have

$$\begin{aligned} f(\boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\rho} \Big| \boldsymbol{\psi}_{-(\eta,\gamma,\rho)}, \boldsymbol{\theta}) \quad &\propto \quad \left\{ \prod_{i=1}^{n} f(I_i\, | Z_i'\boldsymbol{\eta} + \gamma'\boldsymbol{\theta}_i, 1) \right\} \\ &\qquad \left\{ \sum_{j=1}^{\bar{s}} 1(c_{i,j-1} < I_i < c_{i,j}) D_{i,j} \right\} 1(c_{i1} < \cdots < c_{i\bar{s}}) f(\boldsymbol{\eta}, \boldsymbol{\gamma}) f(\boldsymbol{\rho}) \Bigg\}. \end{aligned} \tag{28}$$

This marginal can be factored into two conditionals. Conditional on ρ we

$$f(\boldsymbol{\eta}, \boldsymbol{\gamma}|\boldsymbol{\rho}, \boldsymbol{\psi}_{-(\eta,\gamma,\rho)}, \boldsymbol{\theta}) \propto \prod_{i=1}^{n} f(I_i|Z_i'\boldsymbol{\eta} + \boldsymbol{\gamma}'\boldsymbol{\theta}_i, 1) f(\boldsymbol{\eta}, \boldsymbol{\gamma}).$$

This is the posterior for a normal regression model with covariates $Z_i, \boldsymbol{\theta}_{\mathbf{i}}$ and precision fixed at one. With $f(\eta, \gamma)$ multivariate normal this is a multivariate normal distribution.

The second conditional (for ρ) is

$$f(\boldsymbol{\rho}|\boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\psi}_{-(\eta,\gamma,\rho)}, \boldsymbol{\theta}) \propto \prod_{i=1}^{n} \sum_{j=1}^{\bar{s}} 1(c_{i,j-1} < I_i < c_{i,j}) D_{i,j} 1(c_{i1} < \cdots < c_{i\bar{s}}) f(\boldsymbol{\rho})$$

We sample ρ_s one at a time conditional on the $\rho_1, \ldots, \rho_{s-1}, \rho_{s+1}, \ldots, \rho_{\bar{s}}$. The conditional for ρ_s is

$$\begin{aligned} f(\boldsymbol{\rho}_s \Big| \boldsymbol{\rho}_{-s}, \boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\psi}_{-(\boldsymbol{\eta},\boldsymbol{\gamma},\boldsymbol{\rho})}, \boldsymbol{\theta}) \quad &\propto \quad \prod_{i:s_i=s} 1(c_{i,s-1} < I_i < \mathbf{Q}_{is}\boldsymbol{\rho}_s) \\ \times \prod_{i:s_i=s+1} 1(\mathbf{Q}_{i,s}\boldsymbol{\rho}_s \quad &< \quad I_i < c_{is+1}) \prod_{i=1}^{n} 1(c_{is-1} < \mathbf{Q}_{is}\boldsymbol{\rho}_s < c_{is+1}) f(\rho). \end{aligned} \tag{29}$$

As a prior for $\boldsymbol{\rho}$ we choose $f(\boldsymbol{\rho}) = \prod_s \mathrm{U}(-B, B)$ where $B = 1000$, i.e., a uniform distribution with very large support.

Let K_s be the number of elements in Q_s. We sample $\rho_{hs}, h = 1, \ldots, K_s$ one at a time. The conditional for ρ_{hs} is a uniform distribution with boundary points which can be derived from a series of inequalities. Without loss we can assume that Q_{ihs} is positive. From (29) it follows that

$$\rho_{hs} > \max\Big\{\max_{i:s_i=s} \frac{I_i - \tilde{c}_{is}}{Q_{ihs}}, \max_i \frac{c_{is-1} - \tilde{c}_{is}}{Q_{ihs}}, -K\Big\} \equiv \underline{g}_{hs}$$

$$\rho_{hs} < \min\Big\{\min_{i:s_i=s+1} \frac{I_i - \tilde{c}_{is}}{Q_{ihs}}, \min_i \frac{c_{is+1} - \tilde{c}_{is}}{Q_{ihs}}, K\Big\} \equiv \bar{g}_{hs},$$

where $\tilde{c}_{is} = \sum_{j=1, j\neq h}^{K_s} Q_{ijs}\rho_{js}$. Hence ρ_{hs} is uniform with boundaries $(\underline{g}_{hs}, \bar{g}_{hs})$.

Conditional on the factors we have

$$f(I|\boldsymbol{\psi}_{-(\eta,\gamma,\rho)}, \boldsymbol{\theta}) \propto \Big\{\prod_{i=1}^{n} f(I_i|Z_i'\boldsymbol{\eta} + \boldsymbol{\gamma}'\boldsymbol{\theta}_i, 1)\big\{\sum_{j=1}^{\bar{s}} 1\big(c_{i,j-1} < I_i < c_{i,j}\big) D_{i,j}\big\}.$$

This factors into n independent truncated normals,

$$f(I|\boldsymbol{\psi}_{-(\eta,\gamma,\rho)}, \boldsymbol{\theta}) = \prod_{i=1}^{n} \mathrm{TN}_{(c_{i,j-1}, c_{ij})}(I_i|Z_i'\boldsymbol{\eta} + \boldsymbol{\gamma}'\boldsymbol{\theta}_i, 1).$$

So we sample I_i, $i = 1, \ldots, N$, one at a time from truncated normals.

Measurement equations

The continuous measurement equations are of the form

$$M_{i,j} = X_{m,i,j}'\boldsymbol{\beta}_{m,j}^{c} + \boldsymbol{\alpha}_{m,j}^{c}{}'\boldsymbol{\theta}_i + \varepsilon_{m,i,j}^{c}. \tag{30}$$

Given $X_{m,i,j}, \boldsymbol{\theta}_{\mathbf{i}}$ this is a linear regression model. With multivariate normal priors on $(\beta_{m,j}^{c}, \alpha_{m,j}^{c})$ and a gamma prior on the precision of $\varepsilon^c_{m,i,j}$ this is in the form of the standard conjugate Bayesian linear regression model, with a conditional normal distribution for $\beta^c_{m,j}$ given the precision of $\varepsilon^c_{m,i,j}$ and a gamma distribution for the precision conditional on $\beta^c_{m,j}$.

Let the last $m - m_1$ elements of the measurement vector M be binary indices generated as

$$M^d_j = 1(M^{*d}_j \geq 0), \qquad j = m_1 + 1, \ldots, m.$$

The parameters in the binary measurements are samples as above with two exceptions. First, a separate step samples the latent measurements, M^{*d}_j, as

$$M^{*d}_{i,j} \sim \begin{cases} \mathrm{TN}_{(0,\infty)}\big(M^{*d}_{i,j} | X'_{m,i,j}\boldsymbol{\beta}^d_{m,j} + {\boldsymbol{\alpha}^d_{m,j}}'\boldsymbol{\theta}_i, 1\big) & \text{if } M^d_{i,j} = 1, \\ \mathrm{TN}_{(-\infty,0)}\big(M^{*d}_{i,j} | X'_{m,i,j}\boldsymbol{\beta}^d_{m,j} + {\boldsymbol{\alpha}^d_{m,j}}'\boldsymbol{\theta}_i, 1\big) & \text{if } M^d_{i,j} = 0. \end{cases}$$

Second, the precision is not sampled but fixed at one.

Outcome equations

Let $Y_{s,a}$ be the outcome vector at age a with schooling level s. Suppose both employment and wage outcomes are modeled. Let $Y^c_{s,a}$ be the wage outcome and $Y^d_{s,a}$ the employment outcome. Also let $Y^{*;d}_{s,a}$ be the latent employment index. By the factor structure assumption we have

$$f(Y^c_{s,a}, Y^{*,d}_{s,a}|\boldsymbol{\theta}) = f(Y^c_{s,a}|\boldsymbol{\theta})f(Y^{*,d}_{s,a}|\boldsymbol{\theta}),$$

for a person working.

The model for wages is

$$Y^c_{s,a,i} = X'_{1,a,i}\boldsymbol{\beta}^c_{s,a} + {\boldsymbol{\alpha}^c_{s,a}}'\boldsymbol{\theta}_i + \varepsilon^c_{a,s,i},$$

where $\varepsilon^c_{a,s,i} \sim \mathrm{N}(0, \tau^c_{s,a})$. This is in the form of a standard linear regression model under normality and $(\beta^c_{s,a}, \alpha^c_{s,a}, \tau^c_{s,a})$ is sampled as above (using multivariate normal and gamma priors).

We can allow for general state dependence by modeling the latent employment transition indices as

$$Y^{d,*}_{s,a,i} = \begin{cases} X'_{2,a,s,i}\boldsymbol{\beta}^d_{a,s,0} + {\boldsymbol{\alpha}^d_{a,s,0}}'\boldsymbol{\theta}_i + \varepsilon^d_{a,s,i,0}, & \text{if } Y^d_{s,a-1,i} = 0, \\ X'_{2,a,s,i}\boldsymbol{\beta}^d_{a,s,1} + {\boldsymbol{\alpha}^d_{a,s,1}}'\boldsymbol{\theta}_i + \varepsilon^d_{a,s,i,1}, & \text{if } Y^d_{s,a-1,i} = 1, \end{cases}$$

where $\varepsilon^d_{a,s,i,0}$ and $\varepsilon^d_{a,s,i,1}$ are both standard normal.

The conditional of $(\boldsymbol{\beta}_{2,a,s,0}, \boldsymbol{\alpha}_{2,a,s,0})$ and $(\boldsymbol{\beta}_{2,a,s,1}, \boldsymbol{\alpha}_{2,a,s,1})$ is

$$f\big(\boldsymbol{\beta}^d_{a,s,0}, \boldsymbol{\alpha}^d_{a,s,0}|\boldsymbol{\psi}_{-\beta^d_{a,s,0},\alpha^d_{a,s,0}}\big) \propto f(\boldsymbol{\beta}^d_{a,s,0}, \boldsymbol{\alpha}^d_{a,s,0}) \prod_{i:Y^d_{s,a-1,i}=0} f\big(Y^{d,*}_{s,a,i}|X'_{2,a,s,i}\boldsymbol{\beta}^d_{a,s,0} + {\boldsymbol{\alpha}^d_{a,s,0}}'\boldsymbol{\theta}_i, 1\big)$$

$$f\big(\boldsymbol{\beta}^{\mathbf{d}}{}_{a,s,1}, \boldsymbol{\alpha}^{\mathbf{d}}{}_{a,s,1}|\boldsymbol{\psi}_{-\beta^d_{a,s,1},\alpha^d_{a,s,1}}\big) \propto f(\boldsymbol{\beta}^d_{a,s,1}, \boldsymbol{\alpha}^d_{a,s,1}) \prod_{i:Y^d_{s,a-1,i}=1} f\big(Y^{d,*}_{s,a,i}|X'_{2,a,s,i}\boldsymbol{\beta}^d_{a,s,1} + +{\boldsymbol{\alpha}^d_{a,s,1}}'\boldsymbol{\theta}_i, 1\big)$$

Both of these are normal regression models with the precision fixed at one. The latent employment indices are sampled as in the usual binary choice framework (see Albert and Chib (1993)).

Factors

The conditional for $\boldsymbol{\theta}$ factors into n conditionals for $\theta_1, \dots, \theta_n$. To see what the conditional for θ_i is note that all contributions of θ_i originate from linear regression models,

$$\begin{aligned}
I_i - Z_i'\boldsymbol{\eta} &= \boldsymbol{\gamma}'\boldsymbol{\theta}_i + \varepsilon_{I,i}, && \text{(choice model)}\\
M_j - X'_{m,i,j}\boldsymbol{\beta}_{m,j} &= \boldsymbol{\alpha}'_{m,j}\boldsymbol{\theta}_i + \varepsilon_{m,j}, && \text{(measurements)}\\
Y^c_{s,a,i} - X'_{1,a,i}\boldsymbol{\beta}^{\mathbf{c}}{}_{s,a} &= \boldsymbol{\alpha}^{\mathbf{c}\prime}{}_{s,a}\boldsymbol{\theta}_i + \varepsilon^c_{a,s,i}, && \text{(wages)}\\
Y^{d,*}_{2,s,a,i} - X'_{2,a,s,i}\boldsymbol{\beta}^{\mathbf{d}}{}_{a,s,l} &= \boldsymbol{\alpha}^{\mathbf{d}\prime}{}_{a,s,l}\boldsymbol{\theta}_i + \varepsilon^d_{a,s,i,l}, && \text{(employment).}
\end{aligned}$$

This equation system is of the form

$$\hat{Y}_i = \mathbf{A}_i\boldsymbol{\theta}_i + u_i,$$

where $u_i \sim N(0, \boldsymbol{\Sigma}_i)$, where $\boldsymbol{\Sigma}_i$ is a diagonal precision matrix. The conditional posterior for θ_i is then

$$f(\theta_i|\psi) \propto \exp\Big\{-\frac{1}{2}(\hat{Y}_i - \mathbf{A}_i\boldsymbol{\theta}_i)'\boldsymbol{\Sigma}_i(\hat{Y}_i - \mathbf{A}_i\boldsymbol{\theta}_i)\Big\} f(\boldsymbol{\theta}_i),$$

where

$$f(\boldsymbol{\theta}_i) = \prod_{k=1}^{K}\sum_{j=1}^{J_K} p_{k,j}\mathrm{N}\big(\theta_{ik}|\mu_{k,j}, \tau_{k,j}\big).$$

We sample $\theta_{ik}|\{\theta_{ij}\}_{j\neq k}$ one at a time from their respective conditionals which can be shown to be a mixture of normals with updated (data dependent) mixture weights and parameters.

Conditional on the factor vector θ, we have

$$\theta_{ik} \sim \sum_{j=1}^{J_k} p_{\ell,j}\mathrm{N}\big(\theta_{i\ell}|\mu_{\ell,j}, \tau_{\ell,j}\big), \qquad i = 1, \dots, n.$$

Conditional on θ we can treat the factors as known and update the mixture parameters (p_k, μ_k, τ_k). We follow the “group indicator” approach in Diebolt and Robert (1994) and augment the parameter vector by a sequence of latent group indicators defined as $g_i = j$ if a $\theta_{i,j}$ originates from mixture component j. Conditional on the mixture group indicators the mixture parameters are easily sampled and conditional on the mixture parameters the group indicators are simple multinomials. To preserve identification of intercepts we constrain the mixture to have mean zero using the method proposed in Richardson et al., (2000).

The estimation of the structural models in section 7 are done as above with a few modifications. The choice model is a probit so the cut point is $c = 0$, and no ρ parameters are estimated. The cross equation restrictions are imposed as follows. Let $\tilde{\mathbf{Y}}_{\mathbf{i}} = (V_{1,h,i}, V_{2,h,i}, V_{1,c,i}, V_{2,c,i}, V_i)$, i.e., the stacked outcomes under high school and college and the choice index. We can then write the model as

$$\begin{aligned}\tilde{\mathbf{Y}}_{\mathbf{i}} &= W_i\psi + \Gamma\theta_i + \varepsilon_i,\\ {}_i &= X_i\omega + \alpha_{\text{test}}\theta_i + \varepsilon_{\text{test}},\end{aligned}$$

where $\psi = \big\{\{\bar{\delta}_{1,s}, \bar{\delta}_{2,s}, \bar{\beta}_{1,s}, \bar{\beta}_{2,s}\}_s, \delta_P, \gamma\big\}$, and W_i and the loading matrix $\Gamma = \Gamma(\{\bar{\alpha}_{1,s}, \bar{\alpha}_{2,s}\}_s, \alpha_P)$ are defined appropriately. This model is now in the form of the system described above and the required conditionals are derived as above.

Appendix D: Mean Preserving Spread

For the model described in Section 7, assume that $\varepsilon_{a,s}$ are independent and identically normally distributed within each period:

$$\varepsilon_{a,s} \sim N(0, \sigma^2_{s,1}) \text{ for ages 19-29}$$

$$\varepsilon_{a,s} \sim N(0, \sigma^2_{s,2}) \text{ for ages 30-65.}$$

Then:

$$\overline{\varepsilon}_{1,s} \sim N(0, \sum_{a=19}^{29} \frac{\sigma^2_{s,1}}{(1+\rho)^a})$$

$$\overline{\varepsilon}_{2,s} \sim N(0, \sum_{a=30}^{65} \frac{\sigma^2_{s,2}}{(1+\rho)^a}).$$

At each age:

$$\ln Y_{a,s} = \delta_{a,s} + X'\boldsymbol{\beta}_{a,s} + \boldsymbol{\alpha}'_{a,s}\boldsymbol{\theta} + \varepsilon_{a,s} + \eta_{1,s} * \text{experience}_a + \eta_{2,s} * \text{experience}^2_a = \mu_{a,s} + \ \varepsilon_{a,s}$$

where

$$\mu_{a,s} = \delta_{a,s} + X'\boldsymbol{\beta}_{a,s} + \boldsymbol{\alpha}'_{a,s}\boldsymbol{\theta} + \eta_{1,s} * \text{experience}_a + \eta_{2,s} * \text{experience}^2_a$$

then

$$E(Y_{a,s}|X,\theta) = \exp(\mu_{a,s})E[\exp(\varepsilon_{a,s})].$$

We do a mean preserving spread at *each* age a by giving the individual knowledge of $\varepsilon_{a,s}$:

$$E\left(Y_{a,s}|X, \theta, \varepsilon_{a,s}\right) = \exp\left(\mu_{a,s} + \varepsilon_{a,s}\right) = \exp\left(\mu'_{a,s}\right)$$

Then,

$$\exp\left(\mu'_{a,s}\right) = \exp(\mu_{a,s})E[\exp(\varepsilon_{a,s})]$$

Since the $\varepsilon_{a,s}$ are iid we can drop the age subscript on the ε:

$$\exp\left(\mu'_{a,s}\right) = \exp\left(\mu_{a,s}\right) E\left(\exp\left(\varepsilon_s\right)\right)$$

The mean preserving spread is actually a combination of a age by age mean preserving spreads. Finally, compute:

$$\overline{\mu}_{1,s} = \sum_{a=19}^{29} \frac{\mu_{a,s}}{(1+\rho)^a}$$

$$\overline{\mu}_{2,s} = \sum_{a=30}^{65} \frac{\mu_{a,s}}{(1+\rho)^a}$$

$$\overline{\mu}'_{1,s} = \sum_{a=19}^{29} \frac{\mu'_{a,s}}{(1+\rho)^a}$$

$$\overline{\mu}'_{2,s} = \sum_{a=30}^{65} \frac{\mu'_{a,s}}{(1+\rho)^a}.$$

Define

$$\begin{aligned} V &= \overline{\mu}_{1,C} + \overline{\mu}_{2,C} - \overline{\mu}_{1,a} - \overline{\mu}_{2,a} - Z\boldsymbol{\gamma} - \boldsymbol{\alpha}'_p\boldsymbol{\theta} - \varepsilon^p \\ V' &= \overline{\mu}'_{1,C} + \overline{\mu}'_{2,C} - \overline{\mu}'_{1,a} - \overline{\mu}'_{2,a} + Z\boldsymbol{\gamma} - \boldsymbol{\alpha}'_p\boldsymbol{\theta} - \varepsilon^p + \bar{\varepsilon}_{1,C} + \bar{\varepsilon}_{2,C} - \bar{\varepsilon}_{1,a} - \bar{\varepsilon}_{2,a} \end{aligned}$$

The probability of going to college is simply given by

$$\Pr(V > 0)$$

for the first case and for the second case

$$\Pr(V' > 0).$$

The experiment for the case where we remove θ_1 from the information set of the agent, keeping age by age mean earnings constant, is analogous to the one just described.

References

[1] Aakvik, A., J. Heckman and E. Vytlacil, “Training Effects on Employment when the Training Effects are Heterogeneous: An Application to Norwegian Vocational Rehabilitation Programs,” manuscript, University of Chicago, 1999.

[2] _____, “Treatment Effects For Discrete Outcomes when Responses To Treatment Vary Among Observationally Identical Persons: An Application to Norwegian Vocational Rehabilitation Programs,” NBER Working Paper No.TO262, forthcoming in *Journal of Econometrics*, 2003.

[3] Albert, J. and S. Chib, “Bayesian Analysis of Binary and Polychotomous Response Data,” *Journal of the American Statistical Association* 88 (1993): 669-679.

[4] Anderson, T.W. and H. Rubin, “Statistical Inference in Factor Analysis,” in J. Neyman, ed., *Proceedings of Third Berkeley Symposium on Mathematical Statistics and Probability,* University of California Press, 5, 1956, 111-150.

[5] Athey, S and G. Imbens, “Identification and Inference in Nonlinear Difference-In-Differences Models,” NBER Technical Working Paper T0280, 2002.

[6] Ben Akiva, Moshe; Bolduc, Denis; and Walker, Joan (2001). “Specification, Identification and Estimation of the Logit Kernel (or Continuous Mixed Logit Model). Manuscript, Department of Civil Engineering, MIT, February.

[7] Buera, Francisco Javier, “Testable Implications and Identification of Occupational Choice Models,” unpublished manuscript, University of Chicago, 2002.

[8] Bureau of Labor Statistics (2001). *NLS Handbook 2001.* Washington, D.C.: U.S. Department of Labor.

[9] Cameron, S. and J. Heckman, “Son of CTM: The DCPA Approach Based on Discrete Factor Structure Models,” unpublished manuscript, University of Chicago, 1987.

[10] _____, “Life Cycle Schooling and Dynamic Selection Bias,” *Journal of Political Economy* 106(2)(1998), 262-333.

[11] _____, “The Dynamics of Educational Attainment for Blacks, Whites and Hispanics,” *Journal of Political Economy* 109(3) (2001), 455-499.

[12] Carneiro, P., K. Hansen and J. Heckman, ”Removing the Veil of Ignorance in Assessing the Distributional Impacts of Social Policies,” *Swedish Economic Policy Review,* Vol. 8, (2001).

[13] Cawley, J., K. Conneely, J. Heckman and E. Vytlacil., “Cognitive Ability, Wages, and Meritocracy,” in *Intelligence Genes, and Success: Scientists Respond to the Bell Curve*, edited by B. Devlin, S. E. Feinberg, D. Resnick and K. Roeder, (Copernicus: Springer-Verlag, 1997), 179-192.

[14] Chamberlain, G., “Education, Income, and Ability Revisited,” *Journal of Econometrics* v5, n2 (March 1977a): 241-57

[15] _____, “An Instrumental Variable Interpretation of Identification in Variance Components and MIMIC Models,” in Paul Taubman, ed., *Kinometrics: Determinants of Socio-Economic Success Within and Between Families*, Amsterdam: North-Holland, 1977b

[16] Chamberlain, G. and Z. Griliches, "Unobservables with a Variance-Components Structure: Ability, Schooling, and the Economic Success of Brothers," *International Economic Review* v16, n2 (June 1975): 422-49.

[17] Chib, Siddhartha and Hamilton, Barton. "Bayesian Analysis of Cross Section and Clustered Data Treatment Models" *Journal of Econometrics*, (2000), 97, 25-50.

[18] _____. (2002) "Semiparametric Bayes Analysis of Longitudinal Data Treatment Models," *Journal of Econometrics*, 110(1)(2002):67–89.

[19] Cochrane, W.G. and D. Rubin, "Controlling Bias in Observational Studies: a review." *Sankhya* A 35(1973), 417-446.

[20] Cosslett, Stephen R., "Distribution-Free Maximum Likelihood Estimator of the Binary Choice Model," *Econometrica,* v51, n3 (May, 1983): 765-82.

[21] Diebolt, J. and C.P. Robert, "Estimation of Finite Mixture Distributions Through Bayesian Sampling," *Journal of the Royal Statistical Society, Series B*, 56(1994): 363-375.

[22] DiNardo, J., N. M. Fortin and T. Lemieux, "Labor Market Institutions and the Distribution of Wages, 1973-1992: A Semiparametric Approach," *Econometrica*, 64(1996): 1001-1044

[23] Eckstein, Z. and K. Wolpin, "The Specification and Estimation of Dynamics Stochastic Discrete Choice Models: A Survey," *Journal of Human Resources*, 24(1989), 562-598.

[24] _____. (1999). "Dynamic Labour Force Participation of Married Women and Endogenous Work Experience." *Review Economic Studies* 56 (July): 375-90.

[25] Elrod and M. Keane, "A Factor-analytic Probit Model for Representing the Market Structure in Panel Data," *Journal of Marketing Research* 32(1995), 1-16.

[26] Ferguson, T. S., "Bayesian Density Estimation by Mixtures of Normal Distributions," in M. Rizvi, J. Rustagi, and D. Siegmund, eds., *Recent Advances in Statistics,* New York: Academic Press, 1983, 287-302.

[27] Flavin, M., "The Adjustment of Consumption to Changing Expectations about Future Income," *Journal of Political Economy* 89(1981), 974-1009.

[28] Florens, J., M. Mouchart and J. Rolin. *Elements of Bayesian Statistics.* New York : M. Dekker, 1990.

[29] Fréchet, M., "Sur les tableaux de corrélation dont les marges sont donneés," *Annals Université Lyon,* Sect.A, Series 3, 14(1951), 53-77.

[30] Geweke, J., D. Houser and M. Keane, "Simulation based inference for dynamic multinomial choice models", in B.H. Baltaji, ed., *Companion for Theoretical Econometrics*, 2001, Basil Blackwell, London.

[31] Goldberger, A.S., "Structural Equation Methods in the Social Sciences." *Econometrica*, 40(1972), 979-1001.

[32] Hansen, K., J. Heckman, and K. Mullen, "Ordered Discrete Choice Models with Stochastic Shocks," manuscript, University of Chicago, 2001.

[33] _____, "The Effect of Schooling and Ability on Achievement Test Scores, forthcoming in *Journal of Econometrics,* 2003.

[34] Hansen, K., J. Heckman, and S. Navarro, "Nonparametric Identification of Time to Treatment Models and The Joint Distributions of Counterfactuals," Unpublished manuscript, University of Chicago, 2003.

[35] Hansen, L., W. Roberds and T. Sargent, "Time Series Implications of Present Value Budget Balance and of Martingale Models of Consumption and Taxes," in L. Hansen and T. Sargent, eds., *Rational Expectations Econometrics.* Boulder, CO: Westview Press, 1991.

[36] Heckman, J.,"Statistical Models for Discrete Panel Data," in C. Manski and D. McFadden, eds., *Structural Analysis of Discrete Data With Econometric Applications.* M.I.T. Press: 1981.

[37] _____, "Varieties of Selection Bias." *American Economic Review* 80(2) (May 1990), 313-18.

[38] _____, "Randomization and Social Policy Evaluation," in *Evaluating Welfare and Training Programs,* edited by Charles F. Manski and Irwin Garfinkel, Cambridge, Mass.: Harvard University Press, 1992.

[39] _____, "Micro Data, Heterogeneity, and the Evaluation of Public Policy: Nobel Lecture," *Journal of Political Economy.* 109(4)(2001), 673-748.

[40] Heckman, J. and B. Honoré, "The Empirical Content of the Roy Model," *Econometrica*, 58(5)(1990),1121-1149.

[41] Heckman, J., R. LaLonde, and J. Smith, "The Economics and Econometrics of Active Labor Market Programs," In O. Ashenfelter and D. Card, eds, *Handbook of Labor Economics.* Volume 3. (Amsterdam: Elsevier, 1999).

[42] Heckman, J., L. Lochner and C. Taber (1998a). "Explaining Rising Wage Inequality: Explorations With A Dynamic General Equilibrium Model of Earnings With Heterogeneous Agents." *Review of Economic Dynamics*, 1:1-58.

[43] _____, (1998b). "General Equilibrium Treatment Effects: A Study of Tuition Policy," *American Economic Review*, 88(2):381-6.

[44] _____, (1998c). "Tax Policy and Human Capital Formation," *American Economic Review*, 88(2):293-297.

[45] _____, (2000). "General Equilibrium Cost Benefit Analysis of Education and Tax Policies," in G. Ranis and L.K. Raut, eds., *Trade, Growth and Development: Essays in Honor of T. N. Srinivasan*, Chapter 14. Amsterdam: Elsevier Science, B.V., 291-393.

[46] Heckman, J. and S. Navarro. "Using Matching, Instrumental Variables and Control Functions to Estimate Economic Choice Models," forthcoming in *Review of Economics and Statistics*, 2003.

[47] Heckman, J., and R. Robb. "Alternative Methods for Evaluation the Impact of Interventions." In *Longitudinal Analysis of Labor Market Data*, J. Heckman and B. Singer, eds. (New York: Cambridge University Press,1985).

[48] _____, (1986). “Alternative Methods for Solving the Problem of Selection Bias in Evaluating the Impact of Treatments on Outcomes,” in *Drawing Inferences from Self-Selected Samples*, H. Wainer, ed. (New York: Springer-Verlag, 1986; reprinted in 2000 by Lawrence Erlbaum Associates).

[49] Heckman, J. and J. Smith, “Assessing the Case for Randomized Evaluation of Social Programs.” in *Measuring Labour Market Measures: Evaluating The Effects of Active Labour Market Policy Initiatives*, K. Jensen and P.K. Madsen, eds. (Copenhagen: Ministry Labour, 1993).

[50] _____, “Evaluating the Welfare State,” in S. Strom, ed., *Econometrics and Economic Theory in the 20th Century: The Ragnar Frisch Centennial*, Econometric Society Monograph Series, (Cambridge: Cambridge University Press, 1998).

[51] Heckman, J., J. Smith, and N. Clements, “Making the Most out of Program Evaluations and Social Experiments: Accounting for Heterogeneity in Program Impacts,” *Review of Economic Studies* 64(1997), 487-535.

[52] Hoeffding, W., “Masstabinvariante Korrelationtheorie,” *Schriften des Mathematischen Instituts und des Instituts für Angewandte Mathematik der Universität Berlin*, 5(1940), 1979-233.

[53] Jöreskog, K., “Structural Equations Models In The Social Sciences: Specification, Estimation and Testing.” In *Applications of Statistics,* edited by P.R. Krishnaih. (Amsterdam: North Holland, 1977), 265-287.

[54] Jöreskog, K.G., and Goldberger, A.S.. “Estimation of a Model with Multiple Indicators and Multiple Causes of a Single Latent Variable,” *Journal of the American Statistical Association*, 70(351)(1975):631-639.

[55] Keane, M., and K. Wolpin, “The Career Decisions of Young Men,” *Journal of Political Economy* 105(3) (June 1997): 473-522.

[56] Kotlarski, Ignacy.“On characterizing the gamma and normal distribution,” *Pacific Journal of Mathematics*, Volume 20 (1967), pp. 69-76.

[57] Manski, Charles F. “Identification of Binary Response Models,” *Journal of the American Statistical Association,* 83(403)(1988), 729-38.

[58] Matzkin, R., “Nonparametric and Distribution-Free Estimation of the Binary Threshold Crossing and the Binary Choice Models,” *Econometrica* v60, n2 (March 1992): 239-70.

[59] _____,. “Nonparametric Identification and Estimation of Polychotomous Choice Models,” *Journal of Econometrics*, 58, 137-68, 1993.

[60] Matzkin, R. and A. Lewbel, “Notes on Single Index Restrictions,” unpublished manuscript, Northwestern University, 2002.

[61] McFadden, D., “Econometric Analysis of Qualitative Response Models,” In *Handbook of Econometrics*, Vol. II, Z. Griliches and M. Intrilligator, eds. (Amsterdam: North Holland, 1984).

[62] Muthen, B., “A General Structural Equation Model With Dichotomous, Ordered Categorical and Continuous Latent Variable Indicators,” *Psychometrika* 49(1984): 115-132.

[63] Olley, S. G. and A. Pakes, “The Dynamics of Productivity in the Telecommunications Equipment Industry,” *Econometrica*, 64(6)(1996), 1263-1297.

[64] Rao, Prakasa B.L.S., *Identifiability in Stochastic Models: Characterization of Probability Distributions*, (Boston: Academic Press, 1992).

[65] Richardson, S., L. Leblond, I. Jaussent and P. J. Green, “Mixture Models in Measurement Error Problems, with Reference to Epidemiological Studies,” Working paper, 2000.

[66] Robert, C.P. and G. Casella, *Monte Carlo Statistical Methods*, (New York: Springer, 1999).

[67] Rosenbaum, P. and D. Rubin, “The Central Role of the Propensity Score in Observational Studies for Causal Effects,” *Biometrika* 70 (April 1983), 41-55.

[68] Roy, A., “Some Thoughts on the Distribution of Earnings,” *Oxford Economic Papers*, 3(1951), 135-146.

[69] Rozanov,Y.A., *Markov Random Fields*, (Berlin: Spring Verlag, 1982).

[70] Sen, Amartya Kumar, *On Economic Inequality*, Oxford, Clarendon Press, 1973.

[71] Willis, R. and S. Rosen, “Education and Self-Selection,” *Journal of Political Economy*, 87(5, part 2)(1979), S7-S36.

Notes

[1] Address: Department of Economics, University of Chicago, 1126 E. 59th Street, IL 60637. E-mail: pmcarnei@midway.uchicago.edu. Pedro Carneiro's research was supported by Fundacao Ciencia and Tecnologia and Fundacao Calouste Gulbenkian.

[2] Address: Kellogg School of Management, Northwestern University, Evanston, Illinois 60657. E-mail: karsten-hansen @northwestern.edu

[3] 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. This research is supported by NSF 97-09-873, SES-0099195, and NICHD-40-4043-000-85-261. Heckman's work was also supported by the American Bar Foundation and the Donner Foundation.

[4] Previous versions of this paper were given at the Midwest Econometrics Group, Chicago, October 2000, Washington University St. Louis, May, 2001, the Nordic Econometrics Meetings, May, 2001 and workshops at Chicago August, 2002 and Stanford, January, 2003. A simple version of this paper is presented in Carneiro, Hansen and Heckman (2001). A version of this paper was presented by Heckman as the Klein Lecture at the University of Pennsylvania, September, 28, 2001 and also at the IFAU conference in Stockholm Sweden, October 2001. We are grateful to all workshop participants. We especially thank Mark Duggan, Orazio Attanasio and Michael Keane for comments on the first draft of this paper. We have benefited from discussions with Ricardo Barros, Richard Blundell, Francisco Buera, Flavio Cunha, Mark Duggan, Lars Hansen, Steven Levitt, Bin Li, Luigi Pistaferri and Sergio Urzua on subsequent drafts. We single out Salvador Navarro and Edward Vytlacil for especially helpful comments. We are grateful to Flavio Cunha and Salvador Navarro for exceptional research assistance and hard work.

[5] See Heckman, Lochner and Taber (1998a, 1998b, 1998c; 2000) for a treatment of general equilibrium policy evaluation.

[6] See Heckman and Smith (1993, 1998) and Heckman, Smith and Clements (1997).

[7] Conditions under which $(M, \tilde{M})$ determine the joint distribution are presented in Rozanov (1982).

[8] See *e.g,* Heckman, Smith, and Clements (1997), or Athey and Imbens, (2002).

[9] Mean or median zero assumptions on (U_0, U_1) are also used.

[10] See their papers for exact conditions. Heckman and Smith (1998) present the most general set of conditions.

[11] Aakvik, Heckman and Vytlacil (1999) present other sets of identifying assumptions.

[12] Heckman and Smith (1998) and Heckman, LaLonde and Smith (1999) discuss conditions under which it is possible to estimate (4).

[13] See Eckstein and Wolpin (1999) and Keane and Wolpin (1997).

[14] See Cameron and Heckman (1998), and Hansen, Heckman and Mullen (2001).

[15] In the case of ties, use the choice with the lowest index.

[16] See Hansen, Heckman, and Navarro (2003) for duration models with general forms of dependence functions generated by this type of model.

[17] Strictly speaking, matching models do not distinguish X and Z. See Heckman and Navarro (2003).

[18] See Hansen, Heckman and Mullen (2001) for a comparison among alternative models of completed schooling. Hansen, Heckman and Mullen (2003) develop a parallel analysis for a one factor-multinomial choice model.

[19] Specifically, it is assumed for (8) that $\mu_s(Z)$ is concave in s for each Z (Cameron and Heckman, 1998), that

$$e_s - e_{s-1} = \tau \qquad \text{all } s = 2, .., \overline{S}$$

with e_1 as an initial condition, that

$$(*) \qquad \mu_s(Z) - \mu_{s-1}(Z) = \varphi(Z) + c_{s-1}$$

with $\mu_1(Z)$ as an initial condition and that $c_s \geq c_{s-1}$ for all $s = 1, \ldots, \bar{s}$. Changes in utilities across states are independent of s, except for an intercept. Then in (17) $\varepsilon_W = \tau + e_1$. If we set all of the iid components of (9) to zero (the uniquenesses ε_s) we get the ordered probit model. As noted in the text, and developed in Hansen, Heckman and Mullen (2003), we can generalize this model to allow $e_s - e_{s-1} = \tau + \chi_s$ where $\chi_s \geq 0$ is a one sided random variable and still secure identification. The requirement $(*)$ precludes a strict random utility model because preferences are state specific. (The strict random utility model requires that $\mu_s(Z)$ not depend on s but Z can vary across s. See, *e.g.*, Matzkin, 1993).

[20] Write $\nu_s = \sum_{j=2}^{s} \rho_j$, where $\rho_j \perp\!\!\!\perp \rho_{j'} (j \neq j')$, $\rho_j \perp\!\!\!\perp \varepsilon_W$, $\rho_j \perp\!\!\!\perp (Z, Q)$, $\rho_j \geq 0, \varphi(Z) = Z'\boldsymbol{\eta}$. This model is identified under the assumptions in Cameron and Heckman (1998) even without any exclusion restrictions, so Q_s can just include an intercept. The proof is trivial. Normalize $\rho_1 = 0$. From the first choice we compute,

$$\Pr(D_1 = 1 \,|Z) = \Pr(Z'\boldsymbol{\eta} + \varepsilon_W \leq c_1)$$

so we can identify $f(\varepsilon_W)$, and $\boldsymbol{\eta}$ up to scale σ_W, assuming ε_W and the ρ_j have densities with respect to Lebesgue measure and nonvanishing characteristic function in addition to other standard regularity conditions. We suppress the intercept in Z. One cannot distinguish the intercept from c_1. Proceeding to further choices we obtain

$$\begin{aligned} \Pr(D_1 + D_2 = 1|Z) &= \Pr(Z'\boldsymbol{\eta} + \varepsilon_W \leq c_2 + \nu_2) \\ &= \Pr(\varepsilon_W - \nu_2 \leq c_2 - Z'\boldsymbol{\eta}). \end{aligned}$$

Therefore we can identify $f(\varepsilon_W - \nu_2)$ and c_2 up to scale $\sigma_{\varepsilon_W - \nu_2}$. The scale is determined by the first normalization (relative to σ_{ε_W}). We can estimate $\frac{\sigma_{\varepsilon_W - \nu_2}}{\sigma_{\varepsilon_W}} = \left(\frac{\sigma^2_{\varepsilon_W} + \sigma^2_{\nu_2}}{\sigma^2_{\varepsilon_W}} \right)^{1/2}$ by taking the ratio of the normalized $\boldsymbol{\eta}$ from the second choice probability to the normalized ratio of $\boldsymbol{\eta}$ from the first choice probability for any coordinate of $\boldsymbol{\eta}$. Define $\psi(X)$ as the characteristic function of X. From the assumed independence of ε_W and ν_2, $\psi(\varepsilon_W - \nu_2) = \psi(\varepsilon_W)\psi(-\nu_2)$. Therefore we can identify $\dfrac{\psi(\varepsilon_W - \nu_2)}{\psi(\varepsilon_W)} = \psi(-\nu_2)$, and we can determine $f(-\nu_2)$ from the convolution theorem adopting a normalization for σ_W. Proceeding sequentially, we obtain $\Pr(D_1 + D_2 + D_3 + ... + D_k = 1 \,|Z) = \Pr(Z'\boldsymbol{\eta} + \varepsilon_W \leq c_k + \nu_k)$, and can identify c_k and $f(\nu_k)$ up to the normalization given in the first step. From $f(\nu_k)$, we can use deconvolution to identify $f(\rho_j), j = 2, .., \overline{S}$. See Hansen, Heckman and Mullen (2001) for further details and extensions to factor models. Nowhere in this analysis do we use the assumption that Q_s contains regressors.

[21] Other normalizations are possible. All require that there be at least three measurements on each factor, although we can get by with only one dedicated measurement. Consider the following example (due to Salvador Navarro):

Let $L = 5, K = 2$.

$$\begin{aligned} g_1 &= \theta_1 + \varepsilon_1,\ g_2 = \lambda_{21}\theta_1 + \theta_2 + \varepsilon_2 \\ g_3 &= \lambda_{31}\theta_1 + \lambda_{32}\theta_2 + \varepsilon_3,\ g_4 = \lambda_{41}\theta_1 + \lambda_{42}\theta_2 + \varepsilon_4 \\ g_5 &= \lambda_{51}\theta_1 + \lambda_{52}\theta_2 + \varepsilon_5. \end{aligned}$$

Assuming nonvanishing covariances and factor loadings,

$$\lambda_{32} = \frac{COV\left(g_1, g_5\right) COV\left(g_3, g_4\right) - COV\left(g_3, g_5\right) COV\left(g_1, g_4\right)}{COV\left(g_2, g_4\right) COV\left(g_1, g_5\right) - COV\left(g_1, g_4\right) COV\left(g_2, g_5\right)}$$

if $\lambda_{22}\lambda_{42}\lambda_{51} - \lambda_{41}\lambda_{52} \neq 0$.

Then

$$\lambda_{41} = \frac{COV\left(g_3, g_4\right) - COV\left(g_2, g_4\right)\lambda_{32}}{COV\left(g_1, g_3\right) - COV\left(g_1, g_2\right)\lambda_{32}}$$

if $\lambda_{31} \neq \lambda_{32}\lambda_{21}$.

$$\begin{aligned} \lambda_{21} &= \frac{COV\left(g_1, g_2\right)\lambda_{41}}{COV\left(g_1, g_4\right)},\ \lambda_{31} = \frac{COV\left(g_1, g_3\right)\lambda_{41}}{COV\left(g_1, g_4\right)},\ \lambda_{51} = \frac{COV\left(g_1, g_5\right)\lambda_{41}}{COV\left(g_1, g_4\right)} \\ \sigma^2_{\theta_1} &= \frac{COV\left(g_1, g_4\right)}{\lambda_{41}},\ \sigma^2_{\theta_2} = \frac{COV\left(g_2 g_3\right) - \lambda_{21}\lambda_{31}\sigma^2_{\theta_1}}{\lambda_{32}} \\ \lambda_{42} &= \frac{COV\left(g_2, g_4\right) - \lambda_{21}\lambda_{41}\sigma^2_{\theta_1}}{\sigma^2_{\theta_2}},\ \lambda_{52} = \frac{COV\left(g_2, g_5\right) - \lambda_{21}\lambda_{51}\sigma^2_{\theta_1}}{\sigma^2_{\theta_2}}. \end{aligned}$$

[22] In particular, U_m^d, U_s^d are assumed to have a distribution that is absolutely continuous with respect to Lebesgue measure.

[23] Alternatively, we could normalize the medians to be zero.

[24] For a definition of measurable separability, see Florens, Mouchart and Rolin (1990), section 5.2. The key idea is that we can vary each of the coordinates of the vector freely.

[25] This assumption can be relaxed. It only affects certain normalizations.

[26] In the discussion of equation (21) we could have normalized the variances of the $\sigma^2_{\theta_i}, i = 1, ..., K$ to one rather than certain factor loadings, although this is less straightforward and requires the imposition of certain sign restrictions.

[27] To be able to identify λ^s_{21} we need a third measurement on this factor, which we can get from equation $B_1 + 1$. Since there is no equation B_{K+1}, we require that $B_K - B_{K-1} \geq 3$ in order to be able to identify the loadings on θ_K.

[28] In principle, the future $\left(\boldsymbol{\alpha}^1_a, \boldsymbol{\alpha}^0_a\right)$ can be uncertain at the date decisions are made. Assuming that these factor loadings are independent of $\boldsymbol{\theta}$, we can replace these expressions by $E_{\mathcal{I}_\theta}\left(\boldsymbol{\alpha}^1_a\right), E_{\mathcal{I}_\theta}\left(\boldsymbol{\alpha}^0_a\right)$ without affecting the identifiability of the $\left(\boldsymbol{\alpha}^1_a, \boldsymbol{\alpha}^0_a\right)$, provided the conditions of Theorem 4 are met, but it affects the identifiability and interpretation of $\boldsymbol{\alpha}_P$. A more general version of this model would postulate two random variables $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^*$. Agents act on $\boldsymbol{\theta}^*$ while $\boldsymbol{\theta}$ is the true value. It would be interesting to identify the joint distributions of $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^*$ under (*e.g.*) a rational expectations assumption. We leave this for a later occasion.

[29] We set zero earnings to 1 in this paper.

[30] The run time was about 122 minutes on a 1.2Ghz AMD Athlon PC.

[31] The distributions of the factors by schooling level are shown in Figures A1 and A2 on the website.

[32] The coefficient estimates for the model are posted on the website.

[33] If we consider net gains by subtracting costs the difference between college graduates and high school graduates will be even higher because costs are lower for college graduates.

[34] We can also compute gross utility gains as a percentage of the gross utility in high school as:

$$R = \frac{V_{1,c} + V_{2,c}}{V_{1,h} + V_{2,h}} - 1.$$

See Figure A4 on the website.

[35] If the agent knows θ_1, θ_2, $\varepsilon_{\text{college}}$ and $\varepsilon_{\text{highschool}}$ then he faces no uncertainty.

[36] These results are available on request from the authors, and are posted on the website.

[37] See the numbers posted at the website.

[38] We compute the compensation (which can be negative or positive) required by each individual to keep average earnings the same after the uncertainty is reduced. Then we provide the individual with this compensation together with knowledge of $(\bar{\varepsilon}_{1,c}, \bar{\varepsilon}_{2,c}, \bar{\varepsilon}_{1,h}, \bar{\varepsilon}_{2,h})$ and finally we compute the percentage of individuals who would change their schooling decision if they had knowledge of $(\bar{\varepsilon}_{1,c}, \bar{\varepsilon}_{2,c}, \bar{\varepsilon}_{1,h}, \bar{\varepsilon}_{2,h})$ but had the same present value of earnings in each schooling level. We use the procedure described at the end of Section 4 applied to each period to adjust utility for the effects of mean preserving spreads in earnings (see Appendix D).

[39] This comes from a simulation available on request from the authors.

[40] The same result holds when we consider distributions of utilities instead of distributions of lifetime earnings. See Figure A-15 on the website.

[41] We thank Edward Vytlacil for simplifying and clarifying the statements and proofs of all three theorems in this section.

[42] Using a standard definition of identification, a model (F_U, β) is identified iff for any alternative parameters $(F_U^*, \beta^*) \neq (F_U, \beta)$, there exists some $\varepsilon > 0$ such that

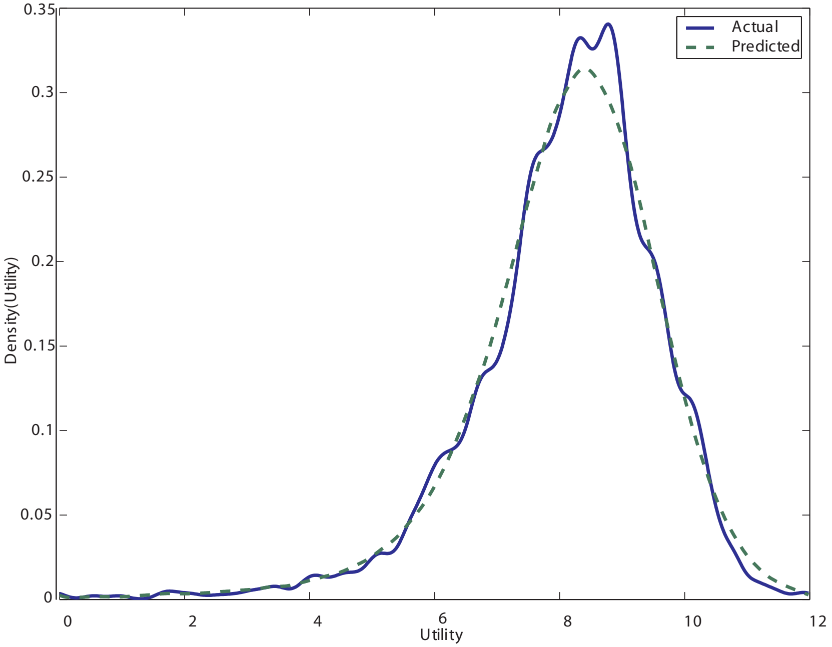
$$\Pr\left(\left|F_U\left(\beta\right) - F_U^*\left(\beta^*\right)\right| > \varepsilon\right) > 0,$$

where the probability is computed with respect to the density of the data generating process. Our use of limit set arguments may appear to contradict the standard definition because of zero probability at the limit sets. However, this intuition is false. See the argument in Aakvik, Heckman and Vytlacil (1999), Theorem 1, which justifies the appeal to limit arguments used in this paper in terms of standard definitions of identification.

[43] Implemented in 1950, the AFQT score is used by the army to screen draftees.

[44] For other uses of Markov Chain Monte Carlo techniques in models and applications related to ours, see Chib and Hamilton (2000) who implements MCMC methods for a panel version of a generalized Roy model and Chib and Hamilton (2002) who consider various cross sectional treatment models.

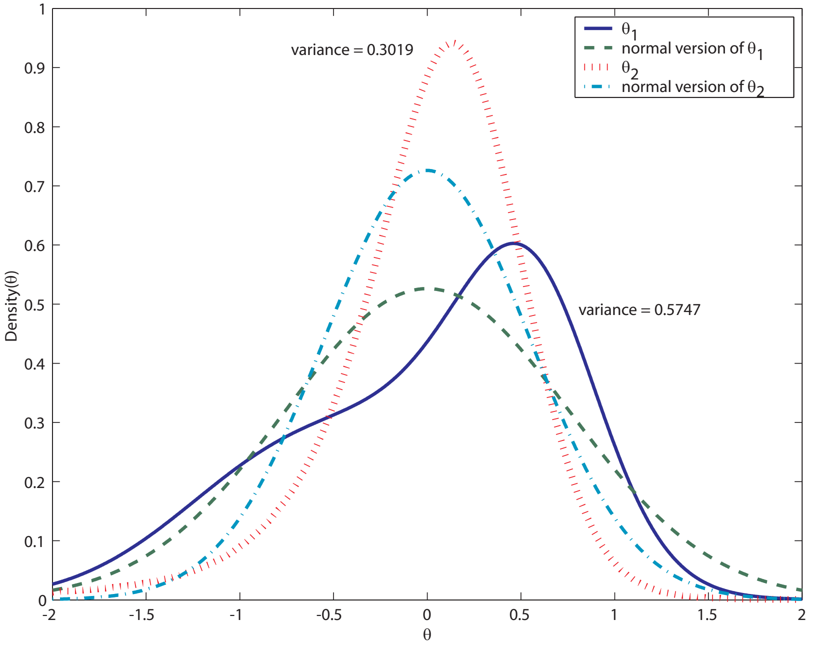
Figure 1
Density of Gross Utility



All densities are estimated using a 100 point grid over the domain and a Gaussian kernel with bandwidth of 0.12.

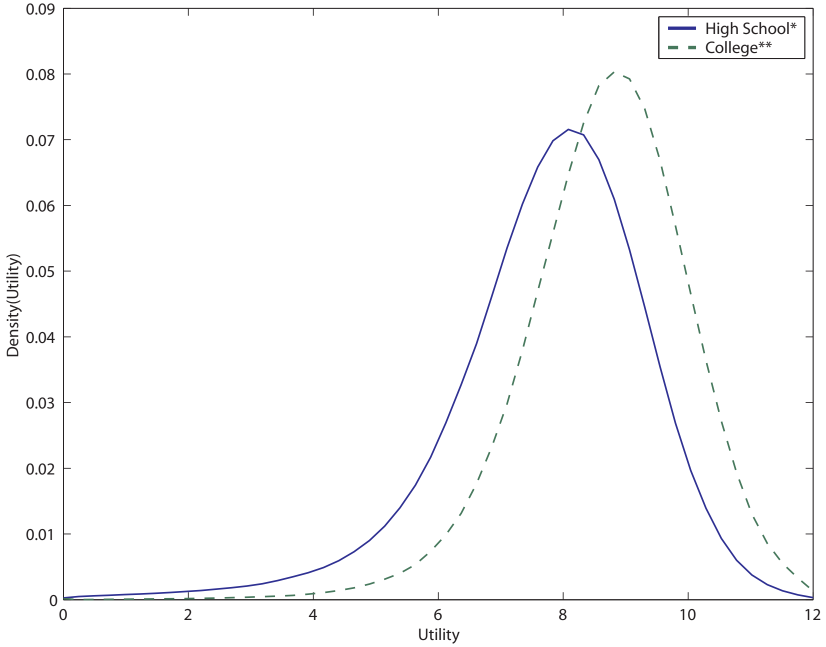
$\text{Utility}=\Sigma_a \frac{1}{(1+0.03)^a}\log(Y_{a,s})$

Figure 2
Factor and Normal Densities*



* Normal densities are defined to be normal with same mean and variance as the corresponding θ. All densities are estimated using a 100 point grid over the domain and a Gaussian kernel with bandwidth of 0.12.

Figure 3
Density of College Gross Utility

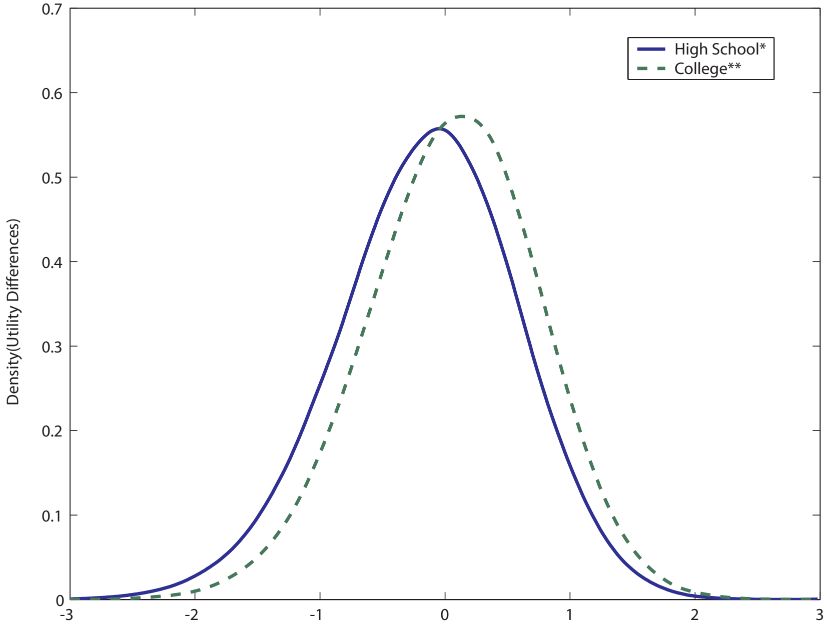


* Counterfactual: E(V_c |C hoice=High S chool)
** Predicted: E(V_c |C hoice=C ollege)

All densities are estimated using a 100 point grid over the domain and a Gaussian kernel with bandwidth of 0.12.

Utility=$\Sigma_a \frac{1}{(1+0.03)^a}\log(Y_{a,s})$

Figure 4
Density of Gross Utility Differences (College-High School)



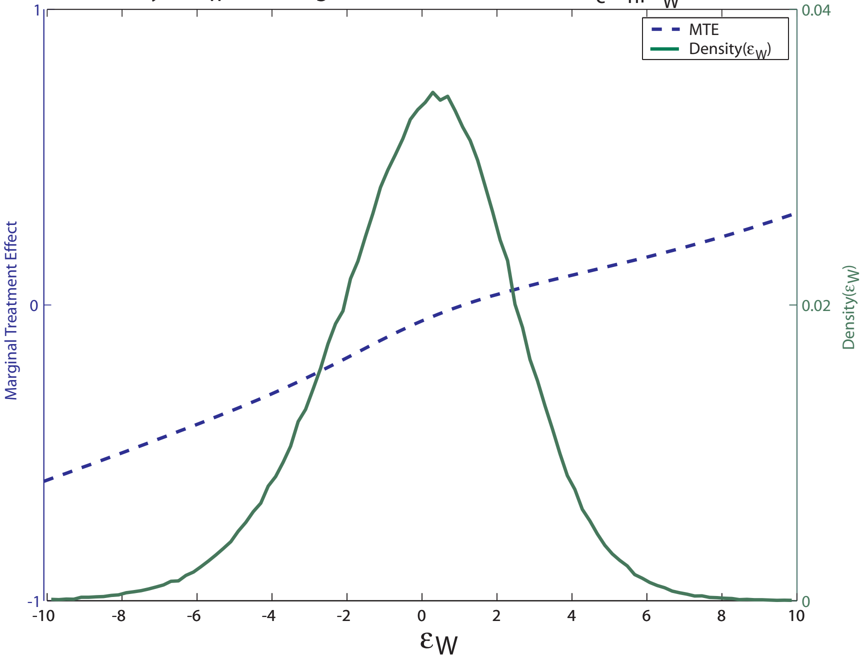
* E(V_c-V_h|Choice=High School)
** E(V_c-V_h|Choice=College)

All densities are estimated using a 100 point grid over the domain and a Gaussian kernel with bandwidth of 0.12.

Utility=$\Sigma_a \frac{1}{(1+0.03)^a}\log(Y_{a,s})$

Figure 5

Density of ε_W and Marginal Treatment Effect: $(E(V_c - V_h | \varepsilon_W))$*

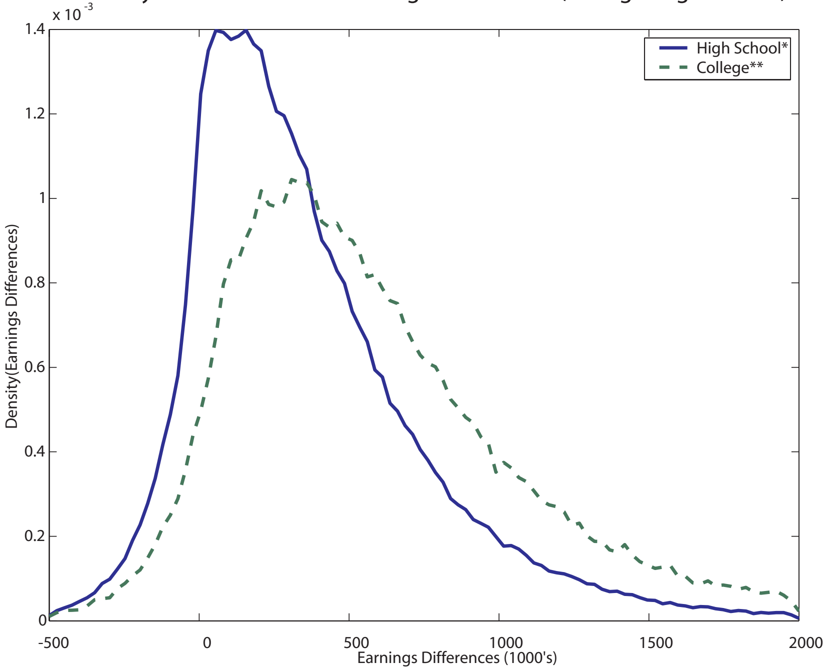


*$\varepsilon_W = (\alpha'_{1,c} + \alpha'_{2,c} - \alpha'_{1,h} - \alpha'_{2,h} - \alpha'_p)\theta - \varepsilon_p$

All densities are estimated using a 100 point grid over the domain and a Gaussian kernel with bandwidth of 0.12.

Figure 6

Density of Gross Lifetime Earnings Differences (College-High School)



*$E(PV_c - PV_h | \text{Choice=High School})$

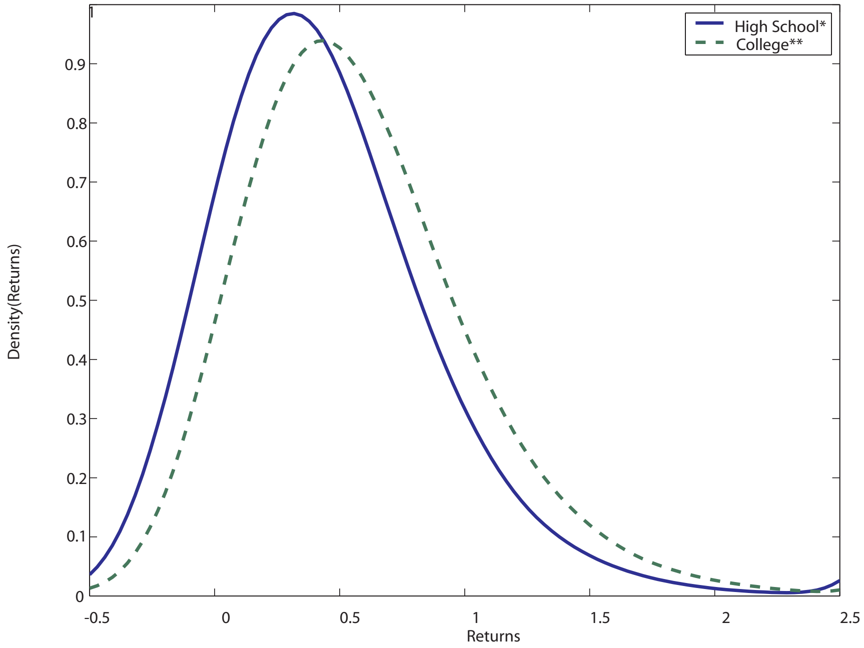
**$E(PV_c - PV_h | \text{Choice=College})$

$PV_h = \Sigma_a \frac{1}{(1+0.03)^a} Y_{h,a}$

All densities are estimated using a 100 point grid over the domain and a Gaussian kernel with bandwidth of 0.12.

Figure 7

Density of Relative Gross Earnings Differences (College-High School)



*E((PV$_c$/PV$_h$)-1|Choice=High School)

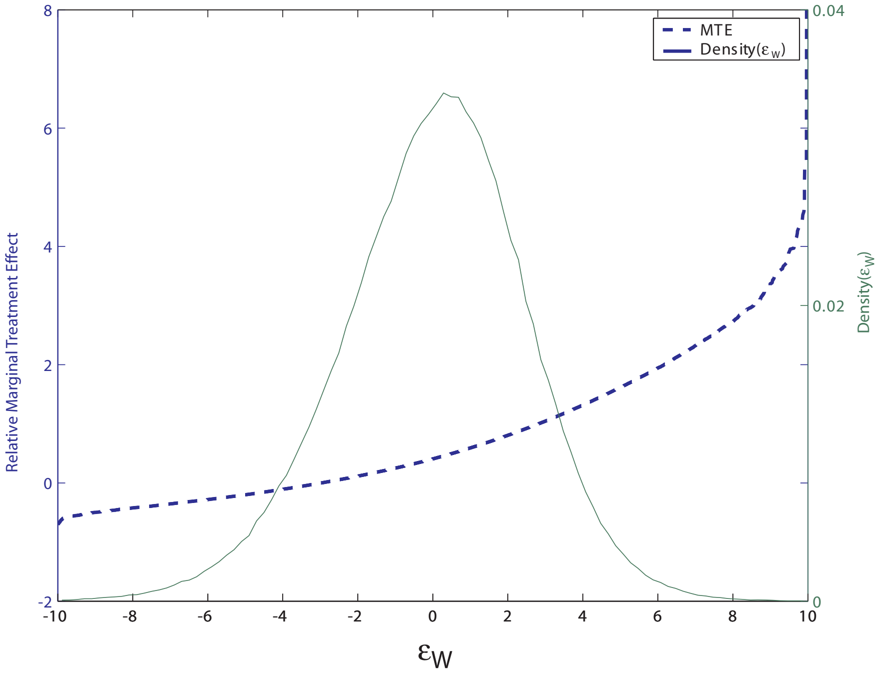
**E((PV$_c$/PV$_h$)-1|Choice=College)

$PV_h=\Sigma_a \frac{1}{(1+0.03)^a} Y_{h,a}$

All densities are estimated using a 100 point grid over the domain and a Gaussian kernel with bandwidth of 0.12.

Figure 8

Density of ε_W and Relative Marginal Treatment Effect for Present Value of Gross Earnings E((PVc/PVh)-1| ε_W)

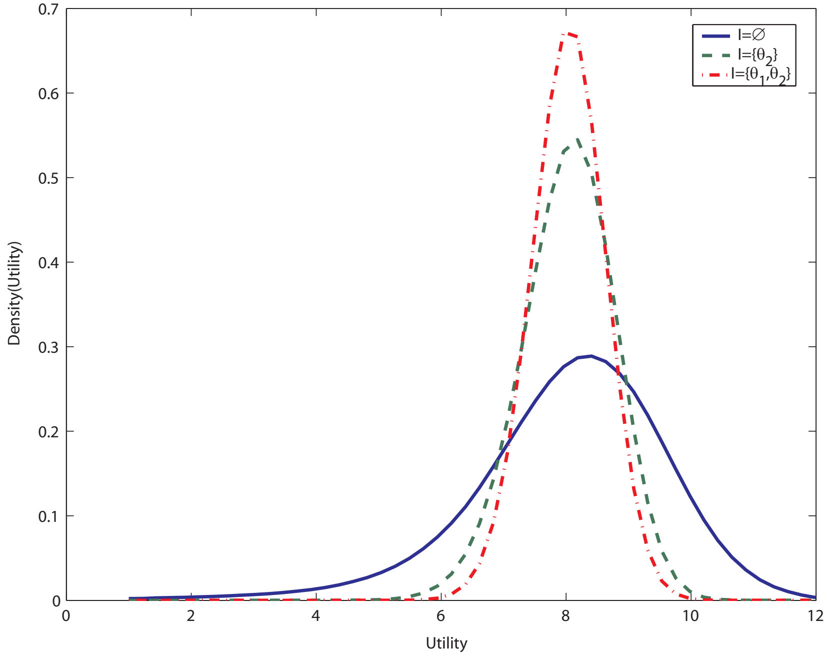


$^*\varepsilon_W = (\alpha'_{1,c} + \alpha'_{2,c} - \alpha'_{1,h} - \alpha'_{2,h} - \alpha'_p)\theta - \varepsilon_p$

All densities are estimated using a 100 point grid over the domain and a Gaussian kernel with bandwidth of 0.12.

Figure 9

Density of Gross College Utility Under Different Information Sets

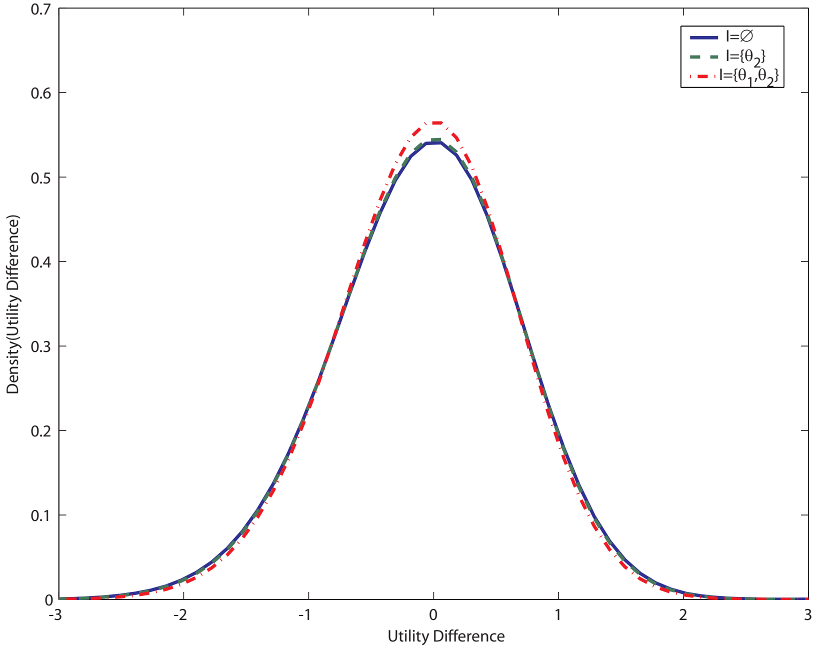


All densities are estimated using a 100 point grid over the domain and a Gaussian kernel with bandwidth of 0.12.

Utility=$\Sigma_a \frac{1}{(1+0.03)^a}\log(Y_{a,s})$

Figure 10

Density of Gross Utility Difference (College-High School) Under Different Information Sets



All densities are estimated using a 100 point grid over the domain and a Gaussian kernel with bandwidth of 0.12.

Utility=$\Sigma_a \frac{1}{(1+0.03)^a}\log(Y_{a,s})$

Figure 11

Proportion of People Induced Into College by Full Subsidy to College Tuition When Information Set = {θ_1, θ_2} by Decile of Initial High School Earnings Distribution

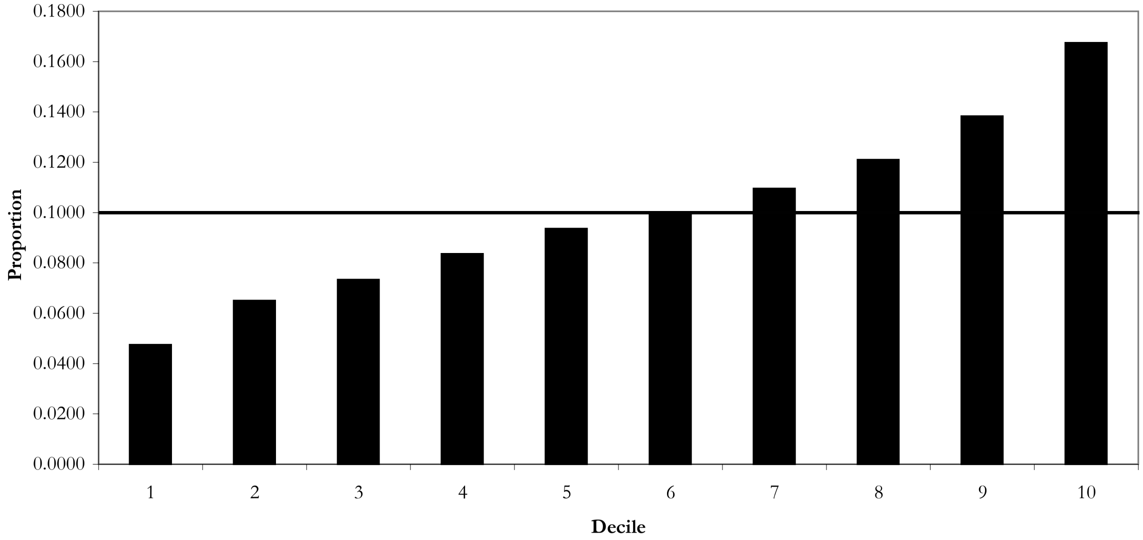


Table 1
Components of $\tilde{G}_s$

	Continuous	Discrete
Variables defined to be the same for all s	M^c	M^d
Variables defined for s	Y_s^c	Y_s^d
Indicator of state	–	D_s

Table 2a
Descriptive Statistics of Variables
NLSY 79 - White Males

	Overall					High School					College				
Variable	Obs	Mean	Std. Dev.	Min	Max	Obs	Mean	Std. Dev.	Min	Max	Obs	Mean	Std. Dev.	Min	Max
Tuition at age 17(Hundreds of Dollars)	1161	20.68	7.70	0.00	53.59	704	21.17	7.85	0.00	47.06	457	19.92	7.40	0.00	53.59
Urban at age 14	1161	0.74	0.44	0.00	1.00	704	0.69	0.46	0.00	1.00	457	0.82	0.38	0.00	1.00
Broken at age 14	1161	0.13	0.34	0.00	1.00	704	0.15	0.36	0.00	1.00	457	0.11	0.31	0.00	1.00
Number of Siblings	1161	2.83	1.77	0.00	15.00	704	3.03	1.85	0.00	15.00	457	2.51	1.60	0.00	11.00
South at age 14	1161	0.19	0.39	0.00	1.00	704	0.19	0.39	0.00	1.00	457	0.19	0.39	0.00	1.00
Mother Education	1161	12.39	2.20	3.00	20.00	704	11.69	1.86	3.00	20.00	457	13.47	2.26	6.00	20.00
Father Education	1161	12.71	3.16	0.00	20.00	704	11.61	2.70	0.00	20.00	457	14.40	3.08	4.00	20.00
Age in 1980	1161	19.27	2.19	16.00	23.00	704	19.27	2.17	16.00	23.00	457	19.28	2.21	16.00	23.00
Distance to College at age 17	1161	7.68	15.58	0.00	100.20	704	8.87	16.01	0.00	100.20	457	5.84	14.75	0.00	96.59
Dummy for Birth in 1957	1161	0.10	0.30	0.00	1.00	704	0.10	0.31	0.00	1.00	457	0.09	0.29	0.00	1.00
Dummy for Birth in 1958	1161	0.10	0.30	0.00	1.00	704	0.09	0.28	0.00	1.00	457	0.12	0.33	0.00	1.00
Dummy for Birth in 1959	1161	0.11	0.31	0.00	1.00	704	0.11	0.31	0.00	1.00	457	0.10	0.30	0.00	1.00
Dummy for Birth in 1960	1161	0.14	0.35	0.00	1.00	704	0.15	0.36	0.00	1.00	457	0.13	0.34	0.00	1.00
Dummy for Birth in 1961	1161	0.14	0.34	0.00	1.00	704	0.14	0.34	0.00	1.00	457	0.13	0.34	0.00	1.00
Dummy for Birth in 1962	1161	0.16	0.37	0.00	1.00	704	0.16	0.37	0.00	1.00	457	0.16	0.36	0.00	1.00
Dummy for Birth in 1963	1161	0.13	0.34	0.00	1.00	704	0.12	0.33	0.00	1.00	457	0.14	0.35	0.00	1.00
Education Status (0 if HS, 1 if College)	1161	0.39	0.49	0.00	1.00	704	0.00	0.00	0.00	0.00	457	1.00	0.00	1.00	1.00
In School at AFQT test date	1161	0.67	0.47	0.00	1.00	704	0.49	0.50	0.00	1.00	457	0.94	0.23	0.00	1.00
Arithmetic Reasoning	1161	0.15	0.95	-2.39	1.42	704	-0.22	0.91	-2.39	1.42	457	0.73	0.70	-1.96	1.42
Word Knowledge (ASVAB 3)	1161	0.14	0.88	-3.71	1.16	704	-0.19	0.92	-3.71	1.16	457	0.64	0.50	-2.24	1.16
Paragraph Composition (ASVAB 4)	1161	0.14	0.89	-3.50	1.21	704	-0.17	0.96	-3.50	1.21	457	0.62	0.47	-1.62	1.21
Coding Speed (ASVAB 6)	1161	0.15	0.95	-3.03	2.79	704	-0.12	0.90	-2.89	2.09	457	0.57	0.87	-3.03	2.79
Math Knowledge (ASVAB 7)	1161	0.14	0.97	-2.14	1.58	704	-0.36	0.80	-2.14	1.58	457	0.91	0.68	-1.83	1.58
Present Value of Earnings*	1161	956.13	730.87	18.12	7861.67	704	694.56	321.93	18.12	1885.85	457	1359.07	964.47	77.02	7861.67

* Earnings in Thousands of Dollars

Table 2b
Descriptive Statistics - Present Value of Log Earnings (Discount Rate = 3%)
NLSY 79 - White Males

	Overall					High School					College				
Variable	Obs	Mean	Std. Dev.	Min	Max	Obs	Mean	Std. Dev.	Min	Max	Obs	Mean	Std. Dev.	Min	Max
Present Value of Earnings (Working Life)*	1161	956.13	730.87	18.12	7861.67	704	694.56	321.93	18.12	1885.85	457	1359.07	964.47	77.02	7861.67
Present Value of Earnings in Period 1*	1161	157.25	72.02	3.91	509.21	704	162.46	74.66	3.91	457.00	457	149.22	67.04	13.55	509.21
Present Value of Earnings in Period 2*	1161	798.88	694.55	14.21	7533.31	704	532.10	251.57	14.21	1582.89	457	1209.84	922.21	63.48	7533.31

*Earnings in Thousands of Dollars
Working Life = From age 19 to age 65
Period 1 = From age 19 to age 29, inclusive
Period 2 = From age 30 to age 65

Table 2c
Covariates Included in Outcome, Choice and Test Equations

	Utility of Earnings	Utility Cost of Schooling	Test Scores
Intercept	Yes	Yes	Yes
Urban	Yes	Yes	Yes
South	Yes	Yes	Yes
Cohort Dummies	Yes	Yes	-
Mean Local Unemployment Rate	Yes	-	-
Average Local Wage	Yes	-	-
Local Tuition	-	Yes	-
Number of Siblings	-	Yes	Yes
Mother's Education	-	Yes	Yes
Father's Education	-	Yes	Yes
Broken Family	-	-	Yes
Enrolled in School at Test Date	-	-	Yes
Age in 1980	-	-	Yes

Table 3a

Factor Loadings

Post-School Utility

		Factor Loading	Standard Error	
Potential First Period	θ_1	0.1419	0.0324	
Utility in High School	θ_2	1	0	
		Total Variance	Proportion of Total Variance Explained by	
		0.3460	θ_1	θ_2
			0.0351	0.8717
Potential Second Period	θ_1	0.2277	0.0519	
Utility in High School	θ_2	1.6432	0.0262	
		Total Variance	Proportion of Total Variance Explained by	
		0.8951	θ_1	θ_2
			0.0349	0.8717
Potential First Period	θ_1	0.1888	0.0559	
Utility in College	θ_2	0.9402	0.0676	
		Total Variance	Proportion of Total Variance Explained by	
		0.3455	θ_1	θ_2
			0.0634	0.7718
Potential Second Period	θ_1	0.3908	0.0979	
Utility in College	θ_2	1.7217	0.1203	
		Total Variance	Proportion of Total Variance Explained by	
		1.0860	θ_1	θ_1
			0.0848	0.8241

Total variance for schooling s in period a is $\alpha^2_{s,a,1}\sigma^2_{\theta_1}+\alpha^2_{s,a,2}\sigma^2_{\theta_2}+\sigma^2_{\varepsilon_{s,a}}$.

Proportion of total variance explained by factor k, in schooling s in period a is $\frac{\alpha^2_{s,a,k}\sigma^2_{\theta_k}}{\text{Total Variance}}$.

Gross Returns

		Factor Loading	Standard Error	
$U_c - U_h$	θ_1	0.2099	0.1553	
	θ_2	0.0188	0.1786	
		Total Variance	Proportion of Total Variance Explained by	
		0.1031	θ_1	θ_1
			0.3027	0.0889

Total variance $= (\alpha_{c,2,1}+\alpha_{c,1,1}-\alpha_{h,2,1}-\alpha_{h,1,1})^2\sigma^2_{\theta_1}+(\alpha_{c,2,2}+\alpha_{c,1,2}-\alpha_{h,2,2}-\alpha_{h,1,2})^2\sigma^2_{\theta_2}+\sigma^2_{\varepsilon_{c,2}}+\sigma^2_{\varepsilon_{c,1}}-\sigma^2_{\varepsilon_{h,2}}-\sigma^2_{\varepsilon_{h,1}}$.

Proportion of total variance explained by factor $k = \frac{(\alpha_{c,2,k}+\alpha_{c,1,k}-\alpha_{h,2,k}-\alpha_{h,1,k})^2\sigma^2_{\theta_k}}{\text{Total Variance}}$.

Table 3b

Factor Loadings

AFQT

		Factor Loading	Standard Error
Arithmetic Reasoning	θ_1	1	0
		Total Variance	Proportion of Total Variance Explained by θ_1
		0.7764	0.7391
Coding Speed	θ_1	0.9672	0.0275
		Total Variance	Proportion of Total Variance Explained by θ_1
		0.7340	0.7308
Math Knowledge	θ_1	0.6313	0.0350
		Total Variance	Proportion of Total Variance Explained by θ_1
		0.8049	0.2843
Word Knowledge	θ_1	0.7508	0.0317
		Total Variance	Proportion of Total Variance Explained by θ_1
		0.6193	0.5219
Paragraph Composition	θ_1	0.8080	0.0345
		Total Variance	Proportion of Total Variance Explained by θ_1
		0.7061	0.5301

Total variance for test t is $\alpha^2_{t,1}\sigma^2_{\theta_1}+\alpha^2_{t,2}\sigma^2_{\theta_2}+\sigma^2_{\varepsilon_t}$.

Proportion of total variance explained by factor k is $\frac{\alpha^2_{t,k}\sigma^2_{\theta_k}}{\text{Total Variance}}$.

Choice

		Factor Loading	Standard Error	
Cost Function*	θ_1	-2.1250	0.5042	
	θ_2	-1.0278	0.3799	
		Total Variance	Proportion of Total Variance Explained by	
		0.8951	θ_1	θ_1
			0.0349	0.9096
Choice**	θ_1	2.3349	0.4904	
	θ_2	1.0466	0.4277	
		Total Variance	Proportion of Total Variance Explained by	
		6.1544	θ_1	θ_1
			0.5297	0.0604

* Cost $= \mu_p+\alpha_{p,1}\theta_1+\alpha_{p,2}\theta_2+\varepsilon_p$.

** Choice $= \mu_{c,2}+\mu_{c,1}-\mu_{h,2}-\mu_{h,1}+(\alpha_{c,2,1}+\alpha_{c,1,1}-\alpha_{h,2,1}-\alpha_{h,1,1}-\alpha_{p,1})\theta_1+(\alpha_{p,2,2}+\alpha_{p,1,2}-\alpha_{h,2,2}-\alpha_{h,1,2}-\alpha_{p,2})\theta_2-\mu_p-\varepsilon_p$.

Total variance of cost $= \alpha^2_{p,1}\sigma^2_{\theta_1} + \alpha^2_{p,2}\sigma^2_{\theta_2} + \sigma^2_{\varepsilon_p}$.

Proportion of total variance of cost explained by factor $k = \frac{\alpha^2_{p.k}\sigma^2_{\theta_k}}{\text{Total Variance of Cost}}$.

Total variance of choice $= (\alpha_{c,2,1}+\alpha_{c,1,1}-\alpha_{h,2,1}-\alpha_{h,1,1}-\alpha_{p,1})^2\sigma^2_{\theta_1}+(\alpha_{c,2,2}+\alpha_{c,1,2}-\alpha_{h,2,2}-\alpha_{h,1,2}-\alpha_{c,2})^2\sigma^2_{\theta_2}+\sigma^2_{\varepsilon_p}$.

Proportion of total variance explained by factor k is $\frac{(\alpha_{c,2,k}+\alpha_{c,1,k}-\alpha_{h,2,k}-\alpha_{h,1,k}-\alpha_{p,k})^2\sigma^2_{\theta_k}}{\text{Total Variance of Choice}}$.

Table 4
Average Gross Utility In Different States (Factual or Counterfactual) For Persons Who Go To High School or Who Go to College and for People At The Margin

(Does Not Include Utility "Cost" or Pyschic Returns to College)

Factual or Counterfactual Schooling Level	High School^{1}	College^{2}	Utility for People at Margin^{3}
High School^{+}	7.8580	8.6125	8.2991
Std. Error	0.0604	0.0737	0.1363
College^{++}	7.7262	8.6885	8.3118
Std. Error	0.0638	0.0763	0.1413

1^{+} E(V$_h$ | choice=high school) and 1^{++} E(V$_c$)|choice=high school)
2^{+} E(V$_h$ | choice=college) and 2^{++} E(V$_c$ |choice=college)
3^{+} E(V$_h$ | V=0) and 3^{++} E(V$_c$ |V=0)

Table 5
Factual and Counterfactual Returns for Persons Who Go To High School, College, or Are At The Margin

(Does Not Include Utility Cost In College)

	High School^{1}	College^{2}	Utility for People at Margin^{3}
Gross Return:			
College Vs. High School (Relative)$^{+}$	-0.0180	0.0126	0.0059
Std. Error	0.1590	0.0178	0.0227
Net Returns:			
College Vs. High School (Relative)$^{++}$	-0.2398	0.3161	-0.0402
Std. Error	0.2502	0.3178	0.0077
College Vs. High School (Relative)$^{+++}$	-0.4227	0.1892	-0.0416
Std. Error	0.5770	0.0144	0.0229

1^{+} E((V$_c$/V$_h$)-1|choice=high school)

2^{+} E((V$_c$/V$_h$)-1|choice=college)

3^{+} E((V$_c$/V$_h$)-1|V=0)

1^{++} E((V$_c$-V$_h$-p)/(V$_h$+p)|choice=high school)

2^{++} E((V$_c$-V$_h$-p)/(V$_h$+p)|choice=college)

3^{++} E((V$_c$-V$_h$-p)/(V$_h$+p)|V=0)

1^{+++} E((V$_c$-V$_h$-p)/(V$_h$)|choice=high school)

2^{+++} E((V$_c$-V$_h$-p)/(V$_h$)|choice=college)

3^{+++} E((V$_c$-V$_h$-p)/(V$_h$)|V=0)

We make the distinction between the second and third line in this table because in our framework we cannot separate nonmonetary costs from nonmonetary benefits of going to college, so we allocate *In* P both ways.

Table 6
Returns to College In Terms of
Lifetime Earnings Excluding Tuition For People
Who Go To H.S., College, or Are At The Margin

	High School[1]	College[2]	Earnings for People at Margin[3]
Gross Returns:			
College vs. High School+	0.4379	0.5764	0.5274
Std. Error	0.0228	0.0365	0.0634
Net Returns:			
College vs. High School++	0.4162	0.5607	0.5092
Std. Error	0.0213	0.0366	0.0605

1^+ E((PV_c/PV_h)-1|choice=high school)

2^+ E((PV_c/PV_h)-1|choice=college)

3^+ E((PV_c/PV_h)-1|V=0)

1^{++} E(($PV_c/(PV_h+PV_{tuition})$)-1|choice=high school)

2^{++} E(($PV_c/(PV_h+PV_{tuition})$)-1|choice=college)

3^{++} E(($PV_c/(PV_h+PV_{tuition})$)-1|V=0)

$PV_j=\sum_a (1/(1+0.03))^a Y_{a,j}$, that is, the interest rate is 3%

Earnings are measured in \$1000s

Table 7
Percentage of People with Negative
Returns to College (Net and Gross)

	Gross		Net*	
	Utility	Earnings	Utility	Earnings
High School Graduates	56.22%	13.62%	95.91%	14.74%
College Graduates	39.66%	6.90%	8.32%	7.28%

* Net means net of total cost for utility and net of tuition costs for earnings.

Table 8

$Pr(d_i < V_c \leq d_{i+1} | d_j < V_h \leq d_{j+1})$ where d_i is the ith decile of the college earnings distribution and d_j is the jth decile of the high school earnings distribution*

High School	College Deciles									
Deciles	1	2	3	4	5	6	7	8	9	10
1	0.7436	0.1936	0.0459	0.0121	0.0035	0.0009	0.0003	0.0001	0.0000	0.0000
2	0.1846	0.3799	0.2372	0.1173	0.0503	0.0206	0.0072	0.0022	0.0006	0.0001
3	0.0482	0.2219	0.2640	0.2078	0.1337	0.0727	0.0344	0.0131	0.0036	0.0005
4	0.0154	0.1108	0.1944	0.2172	0.1902	0.1371	0.0806	0.0389	0.0134	0.0021
5	0.0055	0.0535	0.1240	0.1781	0.1986	0.1807	0.1372	0.0819	0.0341	0.0065
6	0.0019	0.0253	0.0732	0.1274	0.1706	0.1917	0.1826	0.1359	0.0740	0.0175
7	0.0006	0.0103	0.0382	0.0788	0.1271	0.1728	0.2011	0.1926	0.1357	0.0427
8	0.0001	0.0038	0.0171	0.0422	0.0802	0.1300	0.1816	0.2257	0.2173	0.1020
9	0.0000	0.0008	0.0053	0.0165	0.0379	0.0740	0.1288	0.2082	0.2919	0.2365
10	0.0000	0.0000	0.0006	0.0026	0.0079	0.0194	0.0465	0.1015	0.2294	0.5921

*Thus the number in row j column i is the probability that a person with potential high school earnings in the jth decile of the high school earnings distribution has potential college earnings in the ith decile of the college earnings distribution.

Table 9

Agent's Forecast Variance of High-School-College Returns Under Different Information Sets for the Agents

Gross Utility

	$Variance(V_c-V_h)$	$Variance(V_c)$	$Variance(V_h)$	$Correlation(V_c,V_h)$
$I = \varnothing$	0.5134	2.6462	2.3714	0.8990
$I = \{\theta_2\}$	0.5036	0.5068	0.2632	0.3648
$I = \{\theta_1, \theta_2\}$	0.4824	0.3020	0.1804	0

Net Utility

	$Variance(V_c-P_c-V_h)$	$Variance(V_c-P_c)$	$Variance(V_h)$	$Correlation(V_c-P_c,V_h)$
$I = \varnothing$	7.9354	12.7911	2.3714	0.6561
$I = \{\theta_2\}$	7.5549	8.6418	0.2632	0.4476
$I = \{\theta_1, \theta_2\}$	4.4763	4.2959	0.1804	0

Gross Present Value of Earnings

	$Variance(Y_c-Y_h)$	$Variance(Y_c)$	$Variance(Y_h)$	$Correlation(Y_c,Y_h)$
$I = \varnothing$	$2.68x10^5$	$7.69x10^5$	$2.70x10^5$	0.8458
$I = \{\theta_2\}$	$1.18x10^5$	$1.30x10^5$	$2.36x10^5$	0.3234
$I = \{\theta_1, \theta_2\}$	$9.74x10^5$	$8.21x10^4$	$1.53x10^4$	0

Net Present Value of Earnings

	$Variance(Y_c-Tuition-Y_h)$	$Variance(Y_c-Tuition)$	$Variance(Y_h)$	$Correlation(Y_c-Tuition,Y_h)$
$I = \varnothing$	$2.68x10^5$	$7.69x10^5$	$2.70x10^5$	0.8458
$I = \{\theta_2\}$	$1.18x10^5$	$1.30x10^5$	$2.36x10^4$	0.3232
$I = \{\theta_1, \theta_2\}$	9.73x104	$8.20x10^4$	$1.53x10^4$	0

Table 10
People who choose differently under different information sets compensating for the change in risk

Original Choice	Fraction that change choice	
	$I=\{\theta_1,\theta_2,\varepsilon_C,\varepsilon_{HS}\}$	$I=\{\theta_1\}$
High School	0.1181	0.1091
College	0.0159	0.0191
Total	0.0866	0.0813