LIFE CYCLE SCHOOLING AND DYNAMIC SELECTION BIAS: MODELS AND EVIDENCE FOR FIVE COHORTS OF AMERICAN MALES

Stephen V. Cameron
James J. Heckman

Working Paper **6385**

NBER WORKING PAPER SERIES

LIFE CYCLE SCHOOLING AND DYNAMIC SELECTION BIAS: MODELS AND EVIDENCE FOR FIVE COHORTS OF AMERICAN MALES

Stephen V. Cameron
James J. Heckman

Working Paper 6385
http://www.nber.org/papers/w6385

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
January 1998

This research was supported by NSF 87-10145, NSF 91-11455, NSF SBR-93-0248 and grants from the Mellon and Spencer Foundations. An early version of this paper was presented at an Institute For Research on Poverty Workshop in June 1990. We are grateful to Shubham Chaudhuri, Bo Honoré, Michael Keane, Tom MaCurdy, Bob Mason, Rosa Matzkin, Derek Neal, Jose Scheinkman, Michael Rothschild, Joe Tracy, Ed Vytlacil and two anonymous and very helpful referees for comments. We also thank Chris Taber for comments and programming assistance. Any opinions expressed are those of the authors and not those of the National Bureau of Economic Research.

Life Cycle Schooling and Dynamic Selection Bias:
Models and Evidence for Five Cohorts
of American Males
Stephen V. Cameron and James J. Heckman
NBER Working Paper No. 6385
January 1998
JEL Nos. I21, I28, C14

ABSTRACT

This paper examines an empirical regularity found in many societies: that family influences on the probability of transiting from one grade level to the next diminish at higher levels of education. We examine the statistical model used to establish the empirical regularity and the intuitive behavioral interpretation often used to rationalize it. We show that the implicit economic model assumes myopia. The intuitive interpretive model is identified only by imposing arbitrary distributional assumptions onto the data. We produce an alternative choice-theoretic model with fewer parameters that rationalizes the same data and is not based on arbitrary distributional assumptions.

Stephen V. Cameron
Department of Economics
Columbia University
420 West 118th Street
New York, NY 10027

James J. Heckman
Department of Economics
University of Chicago
1126 East 59th Street
Chicago, IL 60637
and NBER
jimh@cicero.spc.uchicago.edu

This paper uses economic theory joined with econometric identification analysis to examine an empirical regularity established in the sociology of education. The estimated effects of family background and family resources on the probability of transiting from one grade to the next diminish at higher levels of education. The empirical regularity refers to the coefficients of logistic regression models for the probabilities of grade transitions. This pattern has been found in studies of U.S. data and in studies of a variety of countries at different stages of economic and political development, including Italy, Switzerland, France, Czechoslovakia, Poland, Great Britain, the Netherlands, Germany, Taiwan and Japan (Mare, 1980 and Blossfeld and Shavit, 1993).

What is an economist to make of this empirical regularity? To answer this question, this paper investigates the following five questions. (1) Is there any reason to be interested in the coefficients of logistic regression transition probabilities? (2) Can the logit model of grade transitions be linked to an economic model with dynamic choice by rational agents? (3) Can the intuitive explanation for the regularity provided by the sociologists - that it is a consequence of dynamic selection bias or uncontrolled unobservables - be formally justified? (4) How sensitive is this explanation to the choice of functional forms for transition equations and the distributions of unobservables? (5) Is it possible to devise a robust econometric scheme that corrects for dynamic selection bias but does not rely on arbitrary distributional assumptions?

Briefly, our answers are as follows. (1) Logit regression coefficients, while widely used to interpret data, are of little interest in their own right. As a purely statistical matter, a variety of alternative measures of the effects of variables on educational attainment transition probabilities might be used. When they are, many show a *reversal* of the empirical regularity displayed by logit regression coefficients. The "empirical regularity" depends on the choice of statistic used to summarize the evidence. The issue of which statistic to use can only be settled by an appeal to behavioral theory and cannot be resolved on purely statistical grounds. (2) The logit model of grade transitions implicitly assumes myopia on the part of agents. It is not an attractive interpretive framework for economists because it is difficult to justify on choice-theoretic grounds. (3) It is possible to formally define dynamic selection bias but it does not universally operate in the way hypothesized in the literature on educational attainment. Selective survival of the "fittest" students can lead to *overstatement* in the estimated effects of family background on higher grade transitions. (4) A pattern of declining logit coefficients for higher grade transitions is critically dependent on choices of functional forms for the distribution of unobservables. (5) Under certain conditions, it is possible to devise robust distribution-free econometric methods to control for dynamic selection bias. Unfortunately, those conditions are not satisfied in the data used to produce the empirical regularity that motivates this paper. In that data, the hypothesis of dynamic selection bias or "educational selectivity" cannot be tested within the grade transition framework.

We propose and estimate a simple low-dimensional alternative model based on rational economic decision-making by agents. The model has many fewer parameters than the models typically estimated in the literature on educational attainment. Within it, it is possible to test for dynamic selection bias without invoking arbitrary distributional assumptions. By a variety of measures, our model fits the data as well or better than the schooling-transition model used in the sociology of education. This is surprising because our model contains many fewer parameters. The economic model explains the schooling attainment of five generations of American males born between 1907 and 1965, and provides a vehicle for interpreting the evidence and conducting counterfactual policy simulations. Using this model, we demonstrate that long-run factors give rise to estimated family income-schooling relationships. It is these long-run family factors and not short-run credit constraints that play a decisive role in explaining schooling attainment.

This paper develops in the following way. Section 1 investigates whether or not there is empirical support for the educational selectivity hypothesis. The hypothesis of educational selectivity is formally stated, and its consequences for the estimated coefficients of commonly-used schooling-transition equations are examined. This analysis is of independent interest since we formally analyze omitted variable bias using global methods, not local methods, in a dynamic, nonlinear setting. We examine the data using a robust, semiparametric estimation method and find evidence that supports the dynamic selection or educational selectivity hypothesis.

We then ask whether estimated coefficients from transition equations are the appropriate objects for investigating or defining the declining pattern of influence of family finances and background on transitions to higher grades. Schooling transition probabilities are nonlinear in their parameters. Other measures of the influence of socio-economic variables on schooling transitions, such as estimated average derivatives, do not follow the pattern found in the estimated logit coefficients. Without an economic model in hand, it is difficult to favor one estimated effect over another in interpreting schooling transition data.

Section 2 develops several simple economic models of schooling attainment to interpret the empirical regularity. This analysis casts light on the behavioral assumptions that underlie the statistical models widely used in the literature. We develop a framework within which the hypothesis of educational selectivity can be explicitly stated and tested, and behaviorally meaningful parameters can be defined. This section also considers identification of the economic model. Understanding the sources of identification of an empirical model is essential to determining under what conditions structural parameters can be recovered and which hypotheses can be tested without imposing arbitrary distributional or functional-form assumptions on estimating equations. A major conclusion of this analysis is that the conventionally-used logit assumption is not innocuous and without invoking a distributional assumption, the hypothesis of educational selectivity is not testable on the data previously used to test it.

Section 3 presents estimates of the structural models of Section 2 for five cohorts of men born between 1907 and 1964. Four cohorts are from the Occupational Change in a Generation (OCG) data widely used to study educational attainment, and one is a recent cohort drawn from the National Longitudinal Survey of Youth (NLSY). We find that a stable-parameter, economically-interpretable low-dimensional model explains the data for all five cohorts of males. Using our estimates, we simulate the likely effect of expansion of financial resources available to parents. This policy increases the enrollment of children in college and raises their graduation rates. However, the responsiveness of both enrollment rates and graduation rates to increases in family income is weak and the policy attracts lower-quality students college. Proposals designed to induce more people to participate in schooling that neglect heterogeneity in student ability, such as those advocated by Jorgenson and Ho (1995), overestimate the wage and productivity effects of increases in college attendance. The students attracted into college by such policies are on average less able and hence less productive than those enrolling in college without the policy's encouragement.

1. Empirical Regularities and Dynamic Selection Bias.

This section begins with a survey of the literature on empirical models of schooling. We present the widely-used schooling-transition model that is used to study the influence of family background on schooling choices. Section 1.2 formalizes the educational selectivity or dynamic selection bias hypothesis by introducing omitted individual effects into the model. We demonstrate the consequences of ignoring omitted effects on empirical estimates of the schooling transition model. Section 1.3 reports estimates of several models and concludes that under one definition of the term, educational selectivity or dynamic selection bias is an important feature of the data on schooling attainment. Section 1.4 introduces alternative ways to measure family background effects on schooling transitions. Using these alternative statistics, the educational selectivity hypothesis is called into question. Since no single statistic best-summarizes family background effects, we conclude that in the absence of a well-defined economic model, the hypothesis of educational selectivity or dynamic selection bias is not well defined.

1.1 Statistical Models of Schooling Choice

Much of the previous research on schooling attainment and family background effects has focused either on a single educational transition, such as college entry among high school graduates, or has used linear regressions to measure the relationship between the highest grade completed and family background variables, such as family income, the highest grade completed of the parents, the number of siblings, and so on. In contrast, Bartholomew (1973) conceptualized schooling attainment as a stochastic process. The sequence of grade transition probabilities generate the probability of schooling attainment. Mare (1980) estimated the schooling-transition model and introduced family background characteristics as determinants of transition probabilities (see also Spilerman, 1977). By dividing schooling into stages, the schooling-transition model parcels out differentials in overall schooling attainment into differentials in transition rates at various stages. This model is the most general statistical model of schooling attainment in the literature on the sociology of education and is the starting point for our analysis.

More formally, let s denote grade in school. Let $D_s = 1$ if a person completes grade s and $D_s = 0$ otherwise. Let $X_s = x_s$ denote regressors determining transitions to grade s from grade s-1. The transition probability from grade s-1 to s is

$$(1) \qquad Pr(D_s = 1 | X_s = x_s, D_{s-1}=1) = P_{s-1,s}(x_s) \; .$$

The event $D_{s-1} = 0$ signifies that an individual drops out at grade s-1; therefore, there is no meaningful event corresponding to the outcome $D_s = 1$ and $D_{s-1} = 0$. These probabilities define the schooling-transition model which becomes a fully specified Markov process when an initial condition D_1 is postulated. Previous analysts typically use a logistic transition probability (see Mare, 1980):

$$(2) \qquad P_{s-1,s}\,(x_s) = \frac{\exp\,(x_s\beta_s)}{1 + \exp\,(x_s\beta_s)} \; .$$

There are several distinctive features of the schooling-transition model. First, the state-space of the model consists of levels of schooling. Since time or individual age does not enter into the definition of the state-space, the model is fundamentally atemporal and does not accommodate time-varying or agerelated regressors.[1] Second, the model ignores "dropping in" or return to school among dropouts and assumes that grade progressions are not affected by any intervening activity between transitions. Third, the model makes no allowance for omitted components of X_s that influence decisions in consecutive transitions but are not measured by observing social scientists. Unmeasured components of "ability", "motivation", and family income come readily to mind as candidate omitted variables that influence outcomes at all transitions. Such omitted variables give rise to the problem of educational selectivity, or dynamic selection bias. Finally, the model is only loosely behaviorally motivated, which makes interpretation of its estimates problematic.

How do other models in the empirical literature on schooling attainment compare to the schooling-transition model? Previous researchers have pursued two distinct lines. First, much of the literature has focused on studying one of the many possible schooling transitions in isolation from the other transitions. For example, Willis and Rosen (1979) and Manski and Wise (1983) investigate only the high school graduation to college attendance transition and do not consider the consequences of estimating their models on a select population of high school graduates. Neither study provides reliable estimates for conducting simulations of policies on college attendance that change the quality of persons graduating from high school.

A second body of literature uses linear regressions to study the relationship between total completed schooling (that is, highest grade completed) and family background. For example, Sewell and Hauser (1975) and Featherman and Hauser (1975) consider only completed schooling which is the sum of the D_s:

$$\text{Completed schooling} = \sum_{s=1}^{\bar{S}} D_s$$

where $\bar{S}$ is the highest attainable grade. Letting $X = (X_1, \ldots, X_{\bar{S}})$, the probability of completing S = s years of school is

$$(3a) \qquad Pr(\sum_{j=1}^{\bar{S}} D_j = s \mid X=x) = \left[\prod_{t=1}^{s} P_{t-1,t}(x_t)\right](1-P_{s,s-1}(x_s))$$

where $P_{\bar{S},\bar{S}+1} = 0$. Expected schooling is

$$(3b) \qquad E\left[\sum_{s=1}^{\bar{S}} D_s \mid X=x\right] = \sum_{s=1}^{\bar{S}} s \, Pr\left[\sum_{j=1}^{\bar{S}} D_j = s \mid X=x\right] .$$

From the viewpoint of the schooling-transition model, linear regression approximations to (3a) and (3b) combine the β_s parameters in an uninterpretable fashion. In this respect, the grade transition model which isolates these parameters is a substantial advance over the models that precede it, even though it does not account for unobserved factors that are correlated across transitions.

Table 1 displays the typical pattern of estimated logit coefficients reported in the schooling attainment literature. The model of equation (2) is estimated on a sample of White American males born in 1937-1946.[2] The first column of the table reports estimates of β_s for the probability of completing elementary school. The subsequent columns report estimates of the coefficients for the educational transition probabilities from the preceding grade to the grade indicated in the column heading. Estimated family background effects (family income and parental education) shrink in absolute value and statistical significance as grade levels increase.

1.2 Educational Selectivity (Dynamic Selection Bias) as a Consequence of Unobserved Heterogeneity

Unobserved heterogeneity in the context of models (1) and (2) arises from omitted components of X_s. In principle, two sources can be distinguished: components that are statistically independent across transitions and those that are not. In order to formalize the discussion of omitted-variable bias in a dynamic model, it is helpful to introduce some new notation. Let Θ_s be a scalar. If the regressors are reordered so that $X_s = (X_{so}, X_{su})$, where "o" is attached to the observed portion of X and "u" is attached to the unobserved or omitted portion of X_s, then in the context of (2),

$$\Theta_s = X_{su}\beta_{su}.$$

To economize on notation, for the rest of this paper let $X = (X_{o1},\ldots,X_{o\bar{S}})$ denote the set of observed regressors.

The hypothesis of heterogeneity bias imposes empirical restrictions on a transition model only if it has some finite-dimensional parameters or other restrictions are imposed such as additive separability between observed and unobserved components of the transition function. In a fully nonparametric representation of the transition probabilities (1), there is no meaningful problem of heterogeneity bias. Representation (3a) is the probability of attaining exactly $S = s$ years of schooling conditional on X. Nonparametric estimates of the transition probabilities exactly reproduce the conditional probabilities in the sample and fully summarize the data. Without any restrictions imposed on the transition probabilities, there is no additional role for unobservables to explain or predict this probability.[3]

In the context of model (2), the problem of heterogeneity is well-defined if two additional restrictions are imposed:

(A-1) $$Pr(D_s = 1 \mid D_{s-1} = 1, X_s = x_s, \Theta_s = \theta_s) = G(x_s\beta_s + \theta_s)$$

where G is a distribution function and

(A-2) $$\Theta = (\Theta_1,..,\Theta_{\bar{S}}) \text{ is independent of } X = (X_1,\ldots,X_{\bar{S}}) \ .$$

The random-effects assumption (A-2) is a point of departure for a more general analysis of the dependence between Θ and X which we do not undertake in this paper. The random-effects assumption is not as restrictive as it first appears. It only characterizes the relationship between Θ and X at the initial schooling stage. Conditional on D_s, Θ and X are not independent, as we illustrate below.

In (A-1), one could replace the linearity in parameters model $(x_s\beta_s + \theta_s)$ by $\varphi_s(x_s,\theta_s)$, where φ_s is a known function or a function with special restrictions placed on it.[4] Other generalizations, such as indexing G by G_s, are possible, but are not considered in this paper. In fact, throughout this paper, except where otherwise noted, we maintain the additional assumption of scalar heterogeneity:

(A-3) $\Theta_s = \Theta$ for all s.

In our context, Θ can be thought of as an "ability" or "motivation" component that is common across transitions.

We further assume

(A-4) $G_s = G$ is known up to some finite parameters,

although this is not strictly required.

Under assumptions (A-1) - (A-4), define

$$Pr(D_1 = 1 \mid X_1 = x_1, \Theta = \theta) = P_{0,1}\,(x_1,\theta)$$

as the probability of completing the first grade while

$$Pr(D_j = 1 \mid X_j = x_j, \Theta = \theta,\, D_{j-1} = 1) = P_{j-1,j}(x_j,\theta)$$

is defined as the probability of transiting to grade j, given completion of grade j-1. The distribution of completed schooling, conditional on (x,θ), is a geometric random variable. Let S be a random variable representing total completed years of schooling. Then the probability of achieving exactly s years of schooling is

$$(4) \qquad Pr\,(S = s \mid X=x,\Theta=\theta)$$

$$= \left\{\prod_{j=1}^{s} Pr(D_j = 1 \mid X_j=x_j,\Theta=\theta,D_{j-1} = 1)\right\} \cdot Pr(D_{s+1} = 0 \mid X_{\bar{S}+1}=x_{\bar{s}+1},\Theta=\theta,D_s = 1).$$

In words, the probability of completing s years of schooling is the probability of reaching s (the term in braces) times the probability of not continuing to $s+1$. When $s = \bar{S}$ (the highest attainable grade of schooling), the artificial random variable $D_{\bar{S}+1}$ is defined so that

$$Pr(D_{\bar{S}+1} = 0 | X=x, \Theta=\theta, D_{\bar{S}} = 1) = 1.$$

This convention allows us to simplify the notation.

There is no empirical counterpart to (4) because Θ is not observed. We can only condition on X in the data. By properties of conditional expectations,

(5) $$Pr(S = s \mid X=x) = E_{\theta}\, (Pr(S = s \mid X=x, \Theta=\theta)) = \int_{\underline{\Theta}} Pr(S = s \mid X=x, \Theta=\theta) dF(\theta)$$

where $F(\theta)$ is the distribution of Θ and $\underline{\Theta}$ is the support of Θ.[5]

A parallel analysis applies to the analysis of the random variable defining whether a person achieves *at least* s years of school:

(6) $$Pr(S \geq s \mid X=x) = E_{\theta}(Pr(S \geq s \mid X=x, \Theta=\theta))$$

$$= \int_{\underline{\Theta}} \prod_{j=1}^{s} Pr(D_j = 1 \mid X_j=x_j, \Theta=\theta, D_{j-1}=1)\ dF(\theta).$$

Assuming independent observations, the sample likelihood for schooling transition models is the product of terms like (5) for uncensored observations and terms like (6) for censored (at s) observations. A sample likelihood that ignores the presence of θ is misspecified and gives rise to the problem of heterogeneity bias in estimating parameters.

The problem of heterogeneity bias in the context of estimating dynamic schooling attainment models is the problem of recovering true features of $Pr(S = s \mid X = x, \Theta = \theta)$ from $Pr(S = s \mid X = x)$. "Features" may be explicit parameters or derivatives or finite change values for the conditioning variables. This problem is interesting if one seeks answers to questions about the effects of variations in X that condition on Θ, as are required in constructing policy counterfactuals. One example addressed in Section 3, asks "what is the effect of an increase in family income on the highest grade of schooling attained holding the ability or motivation of a student constant"? Under assumptions (A-1) through (A-4), we answer this question assuming that ability (Θ) is initially distributed independently of X. If the schooling-transition model could be expressed in the form of a linear regression, the question would be trivially and correctly answered from ordinary least squares estimators derived from regressions of S on X provided other standard conditions are satisfied. It is the nonlinearity of conditional probabilities and the progressively-selective sample composition with respect to Θ at higher levels of educational attainment that gives rise to bias from omitting statistically-independent regressors.

Characterizing Coefficient Bias

Because model (2) is nonlinear, average derivatives of the probabilities with respect to X may be more interpretable than coefficient estimates. However, it is both traditional and analytically fruitful to analyze bias in the coefficients especially if they have a specific meaning within a well-posed behavioral model. We proceed by first characterizing how omitted variables affect coefficient estimates of the initial schooling state (such as completion of elementary school) and then proceed to describe how they affect estimates of coefficients at each stage of the schooling process.

We characterize the bias for β_1 when Θ is omitted and a logit is assumed to characterize the transition probability. To simplify the expression we assume that the regressors are constant across transitions $X_1 = X_2 = \ldots = X_{\bar{S}} = X$ and write

$$Pr(D_1 = 1 | X=x, \Theta=\theta) = \frac{\exp\,(x\beta_1 + \theta)}{1 + \exp(x\beta_1 + \theta)} .$$

Most of the empirical studies of grade transition models use a common set of variables to explain all transitions. We use maximum likelihood to estimate a misspecified logit model that ignores θ

$$Pr(D_1 = 1 | X=x) = \frac{\exp(x\gamma_1)}{1 + \exp(x\gamma_1)},$$

where γ_1 represents parameter estimates from the misspecified model.

Assuming random sampling, in large samples

$$(7)\ plim(\hat{\gamma}_1 - \beta_1) = \int_0^1 \left[E_{X,\theta} \left(\frac{\exp(X\beta_1 + [\theta]\lambda)}{[1+\exp(X\beta_1+[\theta]\lambda)]^2} XX' \right) \right]^{-1} \cdot E_{X,\theta} \left[\frac{\exp(X\beta_1+[\theta]\lambda)}{[1+\exp(X\beta_1+[\theta]\lambda)]^2} X\theta \right] d\lambda,$$

where $E_{X,\theta}$ is the expectation with respect to the joint distribution of (X,Θ).[6] In the case of a scalar X, the bias is positive if the second term under the integral is positive. In general, the bias is not signed but is not zero.[7] The independence of X and Θ does not ensure the absence of large sample bias in estimating β_1. Nonlinearity induces a fundamental nonseparability between X and Θ. The appropriate expression for a general class of discrete-choice models is developed in Appendix A.

In a dynamic model, two factors introduce further bias into estimates of the parameters of the higher grade-transition probabilities. First, the distribution of Θ shifts to the right across successive transitions as low Θ persons drop out of school and hence drop out of the sample. Second, Θ and $X\beta_s$ become negatively correlated because low $X\beta_s$ individuals tend to continue schooling only if they also have high Θ. In other words, among individuals from poor family backgrounds only those with high ability continue schooling. Thus for the population that makes the first transition Θ and X are no longer statistically independent after the first transition.

Expression (7) can be used to make statements about coefficient bias in higher transitions, but the density of Θ with which the expectations of the terms on the right hand side is taken now depends on X and the previous history of schooling choices. To derive the required density, assume that Θ has density $f(\theta)$ and use Bayes' theorem. We obtain the probability of Θ conditional on X and $D_j = 1$, $D_{j-1} = 1, \ldots, D_1 = 1$ by dividing the joint density of Θ, X and $(D_j, D_{j-1}, \ldots, D_1)$ by the joint density of X and

$(D_j,\ldots,D_1)$:

(8)
$$f(\theta \mid X=x,D_j = 1,\ldots,D_1 = 1) = \frac{g(x) \prod_{i=1}^{j} Pr(D_i = 1 \mid X=x,\Theta=\theta,D_{i-1} = 1) f(\theta)d\theta}{g(x) \int_{\underline{\Theta}} \prod_{i=1}^{j} Pr(D_i = 1 \mid X=x,z,D_{i-1} = 1) f(z)dz}$$

$$= \frac{f(\theta) \prod_{i=1}^{j} Pr(D_i = 1 \mid X=x,\Theta=\theta,D_{i-1} = 1)}{\int_{\underline{\Theta}} \prod_{i=1}^{j} Pr(D_i = 1 \mid X=x,z,D_{i-1} = 1) f(z)dz},$$

where $g(x)$ is the marginal density of X.

Two remarks shed light on the properties of the conditional distribution of Θ. First, if all values of the β_j are positive, then

$$\lim_{X\to\infty} f(\theta \mid X = x, D_j = 1,\ldots,D_1 = 1) = f(\theta).$$

In other words, there is no selectivity on θ for high values of x. Second, for fixed x, if $Pr(D_i = 1 \mid X = x, \Theta = \theta, D_{i-1} = 1)$ is increasing in θ for all i, then the mass of the distribution of Θ is reallocated to the upper tail relative to the initial distribution. To illustrate, we rewrite (8) in weighted distribution form as

$$f(\theta \mid X = x, D_j = 1,\ldots,D_1 = 1) = f(\theta)\omega(\theta, x),$$

where

$$\omega(\theta,x) = \frac{\prod_{i=1}^{j} Pr(D_i = 1 \mid X_i=x_i,\Theta=\theta,D_{i-1} = 1)}{\int_{\underline{\Theta}} \prod_{i=1}^{j} Pr(D_i = 1 \mid X_i=x_i,z,D_{i-1} = 1) f(z)dz}$$

is increasing in θ. That is, $f(\theta)$ is upweighted for large values of θ and downweighted for small values when the transition probabilities are monotonically increasing in θ.

To characterize the inconsistency in $\hat{\gamma}_i$ for $i > 1$, replace $\hat{\gamma}_1 - \beta_1$ with $\hat{\gamma}_i - \beta_i$ in equation (7) and take expectations with respect to the conditional density of Θ given X and schooling choices through grade

i - 1. The resulting expression informs us that in general we obtain biased estimates of the parameters β_i if we fit a logit and omit Θ as in the standard piecemeal estimation approach that studies one grade transition in isolation from others. Estimated X effects combine structural and compositional effects (effects of X holding Θ constant and the effects of X on the distribution of Θ at each transition). It is necessary to account for Θ and its changing distribution in order to recover the parameters required for counterfactual policy evaluations.

Can omitted variables produce a pattern of declining estimated coefficients?

The nature of the bias and its effect on the distribution of Θ is illustrated for a ten grade logistic version of the schooling-transition model. Figure 1 illustrates how the distribution of Θ changes across grade transitions. By construction, the initial distribution of Θ is independent of X and normally distributed with a mean of 0 and a variance of 1. (See the base of Table 3 for full details of the models being simulated.) The figure shows the probability that Θ is positive plotted against X for each transition. Since the distribution of Θ is symmetric around zero and is independent of X, the probability of Θ being positive is .5 for all values of X and is represented by the solid line in the figure. At higher grades, the probability that "ability" is positive (*i.e.*, is in the upper half of the original distribution) approaches one for low values of X, while the probability remains close to .5 for high values of X (that is, persons who remain in school by virtue of their high X).

Panel A of Table 2 reveals how dynamic selection alters the distributions of X and Θ in the surviving population. The true coefficient values for X and Θ each are 1, so the probability of continuing in school increases with the value of either variable. Looking across grade transitions, the means of both variables rise (columns 1 and 2), their variances fall (columns 4 and 5) and the correlation between X and Θ becomes increasingly negative (column 3).[8] Panel A of Table 3 displays results from a simple simulation experiment to show how omitted variables produce a pattern of declining estimated coefficients across schooling transitions (column 1). Estimated values of β_s are biased toward zero and the bias increases with each transition (the true value of β_s is 1).

However, this pattern is not universal. Replacing a normally-distributed Θ with a binomial Θ in panel B of Table 2 shows that the correlation between X and Θ initially follows the same pattern, but the correlation begins vanishing after transition 4 (column 3).[9] In addition, the variance of Θ converges to zero implying that there is no heterogeneity bias after a large number of transitions. The subset of persons remaining in school converges to a homogenous population of high Θ individuals. Panel B of Table 3 confirms that the bias in $\hat{\beta}_s$ is nonmonotonic, first increasing and then decreasing after transition 4 (column 1).

Contrary to intuitions expressed in the published literature, an omitted variable such as ability does not necessarily explain the near universal pattern of declining coefficient estimates as the number of transitions becomes large. Nevertheless, our analysis shows that under certain circumstances omitted variables can produce a pattern of estimated coefficients that mimics the pattern found in Table 1 and in many other studies of educational attainment.

1.3 Estimates of the Schooling-Transition Model Corrected For Educational Selectivity

Table 1 displays estimates of the parameters of logistic schooling-transition probabilities obtained using the traditional piecemeal approach. Omitted variables are ignored and separate logit equations are estimated for each schooling transition for the 1937-1946 birth cohort of white American males from the OCG data described in Appendix B. Comparable estimates for older cohorts of the OCG data and a recent birth cohort (1957-1964) from the NLSY are displayed in Appendix Table B.2. Sample means are shown in Table B.1. Parameter estimates of the initial state "Complete Elementary" are listed in the left most column of numbers, and parameter estimates of the next five schooling transitions follow in order from "Attend High School" (attend high school given completion of elementary school) up to "Attend 17+"

(attend any graduate school given college completion). Each transition probability depends on seven family background characteristics, which are listed in the leftmost column. The choices of regressors and schooling categories are made to ensure comparability of our analysis with previous research.[10]

Table 4 presents parameter estimates of the same schooling-transition model augmented to account for omitted variables. Econometrically, the likelihood function is a logistic version of equation (6) with a control for omitted variables that follows a nonparametric estimator of the distribution of Θ introduced by Heckman and Singer (1984) and further analyzed by Follman (1985) and Cameron and Taber (1994). In the nonparametric estimation procedure, no functional form assumptions are made on the distribution of the omitted variable, so in this sense the procedure is robust. The maximum likelihood estimator for $F(\theta)$ puts discrete probability on estimated points of support, where the number of points is also estimated. In any finite sample, it is formally equivalent to a model of discrete Θ.

The declining size and statistical significance of estimated coefficients at higher grade levels is evident in Table 1 for the main family background variables (rows 1 to 4). It is less evident for the binary variables indicating Southern Birth and Farm residence at age 16. However, these variables are not clean measures of family background as they also capture later labor market and schooling opportunities that influence schooling continuation. Comparisons of the estimates reported in Tables 1 and 4 reveal little difference in the estimated coefficients for the first and second transitions; in the higher transitions the heterogeneity-corrected estimates are much greater in absolute value and statistical significance than their uncorrected counterparts.[11] For example, uncorrected estimates of "Number of Siblings" (row 1 in each table) decline in size and statistical significance across transitions while corrected estimates rise after the initial transition and remain relatively constant thereafter. Corrected estimates of parental education ("HGC Father" and "HGC Mother") and Family Income either stay relatively constant or decline much less than their uncorrected counterparts. Corrected and uncorrected estimates for three older cohorts of OCG white males (birth years 1927 to 1936, 1917 to 1926 and 1907 to 1916) and the recent cohort of

NLSY white males (born 1957 to 1964) follow very similar patterns. (See Appendix Table B.3.)

A main conclusion of this section is that educational selectivity caused by omitted variables obscures family background effects. Hence, research reporting piecemeal estimates of the schooling process tends to understate the true effects of family background on educational attainment as measured by the coefficients of logistic transition probabilities.

1.4 Interpreting Estimates Of The Schooling Transition Model

What interpretation should be placed on this finding? While comparisons of estimated logit coefficients have preoccupied the attention of many students of educational attainment, such comparisons can be misleading for several reasons. First, the parameters of a discrete-choice model are identified only up to an arbitrary scale; hence, comparisons of an estimated parameter across transitions or across cohorts cannot separate changes in scale from changes in parameter values. Second, the widely-cited empirical regularity that parameter estimates decline across transitions is an artifact of the logit functional-form assumed in studies documenting this pattern.[12] Indeed, uncorrected coefficients from other models, such as linear probability or probit models, do not display the same declining pattern. Finally, logit coefficients, which represent linear changes in the log-odds ratio, are difficult objects to interpret. The universal empirical regularity that has dominated discussions in the educational attainment literature turns out to be an artifact of the choice of the functional form of an estimating equation.

Panel A of Table 5 shows average derivatives with respect to family background effects for the uncorrected estimates displayed in Table 1. Unlike parameter estimates, estimated average derivatives or average elasticities are much easier to interpret, are less sensitive to functional-form assumptions, and are not altered by a rescaling of the parameters. In general, a variety of summary measures for any data set can be constructed. The partial effects or average derivatives (or transition-specific effects) of panel A are given by $\partial Pr(D_s = 1 \mid X = x, D_{s-1})/\partial x_\ell$ in the first expression where x_ℓ is the ℓ^{th} component of x and the same variables are used across all transitions. The average derivatives reported in panel A of Table 5 do not display the same declining pattern of family background effects as estimated coefficients; instead, they follow an inverted U-shape: rising from the first transition, peaking at college entry, and declining for the last two transitions. Average derivative estimates for the other cohorts follow a very similar pattern.[13]

An alternative characterization of the effect of family background on schooling is the change in the *total probability* of achieving a given schooling level. Panel B of Table 5 presents the estimated elasticity of the probability of grade attainment for each grade decomposed into total effects (operating through all transitions) and partial effects (operating through the indicated transition) for the schooling transition model that does not correct for heterogeneity bias. Panel C is a version of Panel B based on estimates that correct for heterogeneity bias.

To understand the decomposition in Panels A and B of this table, differentiate (6) with respect to the ℓ^{th} component of x, x_ℓ. Thus

$$\frac{\partial Pr(S \geq s \mid X=x)}{\partial x_\ell} = \prod_{j=1}^{s} Pr(D_j=1 \mid X=x, D_{j-1} = 1) \left\{ \sum_{j=1}^{s} \frac{\partial Pr(D_j=1 \mid X=x, D_{j-1}=1)/\partial x_\ell}{Pr(D_j=1 \mid X=x, D_{j-1}=1)} \right\}$$

where $\prod_{j=1}^{s} Pr(D_j=1 \mid X=x, D_{j-1}=1)$ is the probability of achieving at least s years of school (see equation 6). In elasticity form,

$$\frac{\partial \ell n Pr(S \geq s \mid X=x)}{\partial \ell n x_\ell} = \sum_{j=1}^{s} x_\ell \frac{\partial Pr(D_j=1 \mid X=x, D_{j-1})/\partial x_\ell}{Pr(D_j=1 \mid X=x, D_{j-1})} .$$

Panel B reports estimates of these expressions. The terms in parentheses are the elasticities for the indicated transition. The cumulated effects are the main entry of each element in Panel B. The terms in parentheses are the contribution of the row to the total sum. The effects of the variables expressed as elasticities of total change in the probability achieve their greatest value at either the "Graduate College" or "Attend 17+" transitions. Average total derivatives, by contrast, tend to peak at "Attend College". (See Cameron and Heckman, 1994, for these tables.)

How do estimates that do not correct for heterogeneity compare to the heterogeneity-corrected estimates? Panel C presents the analogue of the panel B estimates corrected for omitted-variable bias. Even though uncorrected coefficient estimates tend to be smaller in absolute value than their corrected counterparts, uncorrected elasticities sometimes overstate and sometimes understate the corrected elasticities. For example, comparing the corrected and uncorrected partial elasticities of family income (column 2) reveals that uncorrected estimates are larger at the "Attend College" transition but are smaller thereafter. The elasticity of the total change for the uncorrected model is larger than the corrected elasticity. Corrected partial elasticities are substantially larger than their uncorrected counterparts in the last two transitions and are smaller at the "Attend College" stage. (The uncorrected elasticity is perversely negative in the last transition in the uncorrected case but positive in the corrected case). Total effects for the last three transitions are overstated in absolute value by 5 to 55 percent. (Average total derivatives show comparable biases - see Cameron and Heckman, 1994.) Overall, the corrected partial effects (in parentheses) show their largest influence on the college transitions, while the total effects are generally largest in the last two transitions. The same conclusions hold for the other cohorts of OCG males. See Cameron and Heckman (1994) for these results.

Without a behavioral framework in hand, stability or change in estimated statistical parameters is difficult, if not impossible, to interpret. We have demonstrated that the choice of the parameter of interest critically affects the interpretation of evidence about constancy or invariance across transitions. It is thus of considerable interest to learn whether the schooling-transition model can be rationalized by a behaviorally-coherent economic framework, and whether it is possible to test the hypothesis of educational selectivity without invoking arbitrary distributional and functional form assumptions.

2. Behavioral Models for Interpretation

This section presents interpretable rational choice models of schooling. Before considering the implicit behavioral assumptions built into the grade transition model just discussed, we first present as a benchmark a simple economic model of schooling that is consistent with the evidence reported in Section 1. An identification theorem in Section 2.2 demonstrates that the model is nonparametrically identified so that conclusions from it do not hinge on specific distributional assumptions about unobservables. Within the context of this model the educational selectivity hypothesis imposes testable restrictions on the available data. Section 2.3 explores extensions of this simple model that justify the schooling-transition model as an economic model. These conditions require some form of myopia on the part of the agents being studied and cast the schooling transition model in an unfavorable light. In addition, the model is nonparametrically-nonidentified. A major conclusion of Section 2.4 is that it is necessary to impose distributional assumptions or certain restrictions on the distributions that are not intrinsic to the model of schooling to identify the widely-used schooling transition model using the data analyzed in Section 1.

2.1 An Ordered Discrete Choice Model

We now present a simple wealth-maximizing model of schooling that is consistent with the evidence on schooling transitions for five cohorts of American males. Let the direct costs of schooling, given $X = x$, be denoted by $c(s \mid x)$, which is assumed to be weakly convex, increasing in years of schooling, and a function of family resource and background characteristics as well as any external schooling subsidies. The X are assumed to remain the same across schooling transitions. Assume $c(0 \mid x) = 0$. The discounted lifetime return to schooling is $R(s)$. Assume that $R(s)$ is concave and increasing in s and that $R(0) > 0$. As $R(s)$ is the discounted lifetime return to s years of school, $R(s)$ implicitly includes in its definition both postponed and foregone earnings and it also implicitly embeds discount factors.

Optimal schooling for an individual is the solution to the maximization problem:

$$\max_j \{R(j) - c(j \mid x)\}, \quad j = 0,\ldots,\bar{S}. \tag{9}$$

Our assumptions on return and cost functions guarantee that net returns, $R(j) - c(j \mid x)$, are concave in j for all j and positive for at least the initial schooling state (since $R(0) > 0$ and $c(0 \mid x) = 0$). Ignoring ties, the optimal solution for years of schooling will be unique and positive.

We now introduce an omitted variable into the model that affects the ratio of returns to costs. Let the scalar random variable ε represent a person-specific shifter of the return relative to the cost that is observed and acted on by the individual but not necessarily observed by the analyst. We make two assumptions at this point to facilitate estimation. First, we maintain an assumption analogous to (A-2); that is, ε is independent of X. Second we assume that the relative cost function depends on observed and unobserved individual effects in a multiplicatively separable way:

$$c(j \mid x) = c(j)\varphi(x)\varepsilon,$$

where $E(\varepsilon) = 1$, $\varepsilon \geq 0$, and $\varphi(x) \geq 0$.[14]

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

$$R(s) - c(s)\varphi(x)\varepsilon \geq 0,$$

$$R(s) - c(s)\varphi(x)\varepsilon \geq R(s-1) - c(s-1)\varphi(x)\varepsilon,$$

$$R(s) - c(s)\varphi(x)\varepsilon \geq R(s+1) - c(s+1)\varphi(x)\varepsilon.$$

Thus, the value of ε for a person taking s years of schooling satisfies the following inequalities:

$$\frac{R(s) - R(s-1)}{[c(s) - c(s-1)]\varphi(x)} \geq \varepsilon \geq \frac{R(s+1) - R(s)}{[c(s+1) - c(s)]\varphi(x)} \tag{10}$$

and

$$[R(s) \,/\, c(s \mid x)] \geq \varepsilon.$$

The second inequality ensures that net returns are positive at the optimal level of schooling and will always be satisfied at the optimum since net returns for the initial state are positive by assumption. The first inequality comes from the conditions that guarantee that net returns are maximized at s and has a simple interpretation: $R(s) - R(s-1)$ is the marginal return of s years of school and $[c(s) - c(s-1)]\varphi(x)$ is the marginal cost for an individual with endowment x. Hence, ε is bounded by the ratio of marginal return to marginal cost. This justifies our interpretation of ε as affecting the ratio of returns to costs.[15]

Assuming that ε is continuously distributed, the fraction of the population with observable characteristics x who choose s years of schooling is thus

$$Pr(S = s \mid X=x) = Pr\left[\frac{R(s+1)-R(s)}{[c(s+1)-c(s)]\varphi(x)} \leq \varepsilon \leq \frac{R(s)-R(s-1)}{[c(s)-c(s-1)]\varphi(x)}\right] .$$

This expression defines an ordered discrete-choice model and is a basis for inference about the relationship between family background and schooling choices.

Introduce the function $\ell(s)$ to simplify the notation:

$$\exp(\ell(s)) = \frac{R(s+1) - R(s)}{[c(s+1) - c(s)]} , \qquad s = 1,\ldots,\bar{S},$$

so that

$$Pr(S=s \mid X=x) = Pr\left[\frac{\exp\ \ell(s)}{\varphi(x)} \leq \varepsilon < \frac{\exp\ \ell(s-1)}{\varphi(x)}\right] .$$

By further assuming

$$\varphi(x) = \exp(-x\beta),$$

the ordered-discrete choice problem takes a more familiar form.[16] After rearranging terms we obtain

(11) $$Pr(S=s \mid X=x) = \int_{\ell(s)+x\beta}^{\ell(s-1)+x\beta} f(\log\varepsilon)d\varepsilon ,$$

where f is the density of $\log \varepsilon$. This expression defines the probability that an individual with background characteristics x will complete exactly s years of schooling.[17]

2.2 Is the Model Identified?

An important question regarding the empirical content of any model is whether the structural parameters of this model, $(\ell(s), \beta, F)$, can be estimated without making arbitrary assumptions about the distribution of $\psi = \log \varepsilon$. If the model is not identified without invoking such assumptions, results obtained from it are intrinsically suspect. The following theorem shows that $F(\cdot)$ can be identified without invoking arbitrary distributional assumptions, the $\ell(s)$ parameters can be recovered up to an affine transformation, and the β can be recovered up to a scale transformation as in a standard binary-choice model.[18]

Theorem 1: Let F_ψ denote the distribution of $\psi = \log\varepsilon$. From data on schooling choices $\{D_j\}_{j=1}^{\bar{S}}$ and individual characteristics X and under conditions (i)-(iii) stated below, it is possible to identify β, $\underset{\sim}{\ell} = (\ell(1),\ldots,\ell(\bar{S}))$, and F_ψ up to the following affine transformations:

$$\ell(j) = \mathrm{d}k(j) + \mathrm{e}, j = 1,\ldots,\bar{S}$$

$$\beta = \mathrm{d}\gamma$$

$$F_\psi(\mathrm{d}\eta + \mathrm{e}) = F_\eta(\eta),$$

where $\mathrm{d} > 0$ (a scale parameter), e is a location parameter and $(k(j), \gamma, F_\eta)$ are the parameters of a model observationally equivalent to $(\ell(j), \beta, F_\psi)$. The indeterminacy in the parameters can be resolved by assuming that a quantile of ψ is known and by defining all parameters relative to a scale parameter. The X do not include a constant. The conditions for identification are

(i) F_ψ is absolutely continuous with density $f(\psi) > 0$ almost everywhere and the support of $\psi = (L, U)$, where L and U can be $-\infty$ or ∞ respectively.

(ii) ψ is stochastically independent of X and F_ψ is not a function of X.

(iii) X is a K vector excluding the intercept or other constants and $X \in \bar{\bar{X}} \subseteq R^K$. For some $m \in \{1,\ldots,K\}$, coordinate X_m varies continuously over R^1 and the $(K\text{-}1)$ vector X_{-m}, defined as

X excluding X_m, is not restricted to a hyperplane in R^{K-1}.

Proof: See Appendix A. ■

Assumptions (i) and (ii) guarantee a well-defined distribution function for unobservable ψ and the independence of X and ψ.[19] Assumption (iii) ensures that X has full rank and that it contains at least one element that is continuous. A standard result in discrete-choice analysis is that the coefficients of models are identified only up to scale and location. Hence there are parameters $k(j)$ and γ that require normalizations to pin down their values. Standard ordered discrete-choice models recover slopes and intercepts typically by assuming that $E(\psi) = 0$ or that Median $(\psi) = 0$ to tie down the location of the distribution of the unobservable. They also normalize the Var(ψ) to some value to tie down the scale. Thus it is possible to identify the parameters of the ordered-discrete choice model up to affine transformations and to recover the distribution function of ψ up to an affine transformation (that is, up to scale and location parameters). With data on direct costs, or components of direct costs, it is also possible to identify the scale. The latter case is equivalent to the situation where one component of β is known.

2.3 Can an Economic Model Rationalize the Schooling-Transition Model?

The statistical schooling-transition model of Section 1 is not rationalized by the economic model just presented. In order to obtain grade-specific transition probabilities using the economic framework of Section 2.1, it is necessary to introduce grade-specific shocks into the model and to suppress rational forward-looking behavior. One approach that does this envisions decision-makers who receive multiplicative transition-specific shocks υ_s to their cost or return functions for $s = 1,\ldots,\bar{S}$ and who then solve an optimal schooling problem.

Even if the υ_s are known, sequential stopping rules like those presented in Section 2.1 do not necessarily characterize an optimum unless weak convexity in s is assumed. A sequential rule that has an agent continue in school at least through grade j if

(12a) $$R(j) - c(j)\varphi(x)\varepsilon\upsilon_j \geq R(j\text{-}1) - c(j - 1)\varphi(x)\varepsilon\upsilon_{j-1}$$

is not always optimal. Individuals may continue in school even when (12a) is not satisfied. To see this, first consider the case of perfect foresight in which all υ_j are known at the beginning of life. The υ_j act as grade-specific shocks to the cost function. Even if the sequential rule (12a) is not satisfied, a low υ_{j+2} may make it profitable to make the transition from j to $j+1$ in order to reach $j+2$. That is, application of continuation rule (12a) will not in general lead to a global optimum.

Letting future υ_j be uncertain reinforces this point. Consider the following three grade example. Let $V_j(\cdot)$ represent the value of continuing to grade j given the information available at grade j. Upon completing grade 1, an individual observes υ_1 and continues to grade 2 if and only if

(12b) $$E(V_2(\upsilon_2,\upsilon_1)\,|\,\upsilon_1) \geq R(1) - c(1)\varphi(x)\varepsilon\upsilon_1,$$

where $V_2(\upsilon_2,\upsilon_1) = \text{Max}\{R(2) - c(2)\varphi(x)\varepsilon\upsilon_2,\ R(3) - c(3)\varphi(x)\varepsilon E(\upsilon_3 \mid \upsilon_2,\upsilon_1)\}$. A myopic sequential decision rule of the form (12a) is valid for characterizing the schooling continuation decision when $\text{E}(\nu_2 \mid \nu_1)$ replaces ν_2 only when the value of $R(2) - c(2)\varphi(x)\varepsilon\upsilon_2$ is always greater than or equal to $\text{R}(3)\text{-c}(3)\varphi(\text{x})\varepsilon\, E(\nu_3 \mid \nu_2,\nu_1)$ for almost all ν_2 and ν_1, which would imply that no one would attend grade 3. Uncertainty introduces an additional option value for schooling, and individuals may continue schooling in order to take advantage of the possibility of high future returns.[20]

Observe that if ν_3, ν_2, ν_1 are mutually independent, and we approximate the difference in the expected value function and the first grade return $[\text{E}(\text{V}_2(\nu_2,\nu_1) \mid \nu_1) - (\text{R}(1)\text{-c}(1)\varphi(\text{x})\varepsilon\nu_1]$ by a linear in parameters model with logit separable errors, following Heckman (1981), we can justify the logit transition probability model as an approximation to the true linear model. In this approximation, the link between the economic model and the econometric model is obscured, and the hypothesis of constancy of the β_s across transitions, assumed as the benchmark case in the literature on educational selectivity, is difficult to maintain.

What type of error process justifies an independent grade-specific shock and produces the schooling-transition model? If future values of υ_j are uncertain and agents behave as if the ν_j are martingales so $E(\upsilon_{j+i} \mid \upsilon_j) = \upsilon_j$, $i \geq 1$, and decision making is myopic, then using first period information

$$E(V_2(\upsilon_2,\upsilon_3) \mid \upsilon_1) = \text{Max}\{R(2) - c(2)\varphi(x)\varepsilon\, E(\upsilon_2 \mid \upsilon_1),\ R(3) - c(3)\varphi(x)\varepsilon\, E(\upsilon_3 \mid \upsilon_1)\}$$
$$= \text{Max}\ \{R(2) - c(2)\varphi(x)\varepsilon\upsilon_1,\ R(3) - c(3)\varphi(x)\varepsilon\, \upsilon_1\}$$

so (12a) is equivalent to (12b). In this setup, individuals ignore the potential value of future shocks and act as if today's shock determines the value of all future shocks. Such strong assumptions about expectation formation are not attractive. It is certainly not rational for agents to behave in this way if the υ_j are independent, nor is it an attractive feature of the model for them to ignore the possibility of updating information about υ_j in evaluating future returns if the ν_j are not independent.[21] Moreover, this model does not justify the form of the grade-specific unobservable or the invariant unobservables introduced in Section 1. The unobservables would be dependent across grade transitions and would be based on different random variables in different grades as shocks are experienced and cumulated.

If agents are myopic, decision rule 12(a) accurately represents behavior. In this case rewriting (12a) to incorporate regressors as determinants of the cost function, we obtain

$$\text{(12c)} \qquad \frac{R(j) - R(j-1)}{\varphi(x)[c(j|x) - c(j-1|x)]} \geq \varepsilon\upsilon_{j-1}.$$

Letting $\varphi(x) = \exp(-x\beta)$ and $\log \upsilon_{j-1}$ be logistic, we obtain the schooling-transition model with equation (2) if ε is assumed identical for everyone and set to one for convenience:

(12d)
$$Pr(D_{s+1} | D_s = 1, X=x) = \frac{\exp\left[\frac{\ell(s)+x\beta}{\sigma_s}\right]}{1 + \exp\left[\frac{\ell(s)+x\beta}{\sigma_s}\right]}$$

where σ_s^2 = Var [log v_s] and the $\ell(s)$ are the marginal return to marginal cost parameters previously defined. Reintroducing omitted variables ε in the conditioning set along with the X, we obtain the schooling-transition model introduced in Section 1. Conditional on X, ε and D_{s-1}, the conditional probability of attending grade $s+1$, given that a person attended grade s, is letting $\omega = -\log\varepsilon = -\psi$,

(12e)
$$Pr(D_{s+1} | D_s = 1, X=x, \omega) = \frac{\exp\left[\frac{\ell(s)+x\beta+\omega}{\sigma_s}\right]}{1+\exp\left[\frac{\ell(s)+x\beta+\omega}{\sigma_s}\right]} .$$

Letting $\omega = \theta$ in formula (4), the probability of achieving exactly s years of schooling is given by equation (5).

How does this model compare to the statistical model of Section 1? The two are identical except that the model presented in Section 1 allows for transition-specific coefficients, β_s. The model derived here assumes that β is constant across transitions. This assumption is consistent with a cost function that is a stable across grade transitions. A model of myopic decision making with transition-specific costs of schooling that preserves the convexity of the net reward function in terms of s fully rationalizes the school transition model presented in Section 1.

2.4 The Identifiability of Models With Transition-Specific Innovations

The schooling-transition model presented at the end of Section 2.3 assumes myopia on the part of agents. We now demonstrate that the model is not nonparametrically identified using the type of data analyzed in Section 1. Thus, tests of the educational selectivity hypothesis based on those data are critically dependent on choices of functional forms for estimating equations. Using the data analyzed in Section 1, the model has no empirical content apart from the parametric structure imposed on it. What you get from the model is what you put into it.

In the general case, ε cannot be separately identified from the υ_s. Therefore, we begin our discussion by redefining the υ_s to include ε. We work with $\eta_s = \log \upsilon_s + \log \varepsilon$ and examine identification of the $\overline{S}$-dimensional distribution F_η, the $\ell(s)$, and β_s. We do not impose independence or any other assumption on the distribution of the shocks.

What type of restrictions are needed for the model to be identified? Theorems 2 to 5 below inform us that we need restrictions on X, β_s, or the distribution of η_s. The regressor set X must satisfy certain exclusion restrictions such as having some component of X that varies across grades. For example, a measure of family income or wealth that has some independent variation across grades would suffice. Interestingly, Theorem 3 tells us that, even with data in which the X_s are the same across all grades, if β_s assumes transition-specific values then the model can be fully identified as long as no β_s is a scalar duplicate of another one (that is for all j, there are no $k \neq j$ such that the vector $\beta_j = q\beta_k$ for some nonzero constant q and for some $k \neq j$) provided certain regularity assumptions are met. Thus the models of Section 1 can be identified with no functional-form restrictions on the distribution of the error terms, but the common β restriction cannot be tested.

Theorems 4 and 5 state restrictions on the distribution of $\eta_1,..,\eta_{\overline{S}}$ that suffice to identify the schooling transition model even when β and X are the same across schooling transitions. In these theorems, we assume that the η_s are conditionally independent given common component ν and assume that the innovations come from a known distribution, such as a logistic. The distribution of the heterogeneity component, ν, is nonparametrically identified. These results establish identification of the estimator used to produce the estimates reported in Section 1. We now present the first theorem for the schooling transition model.

Theorem 2. For the schooling model (12a) with $\log c(s \mid x) = \log c(s) - x_s\beta_s + \eta_s$, $s = 1,\ldots,\bar{S}$, assume that distribution of the transition-specific shocks $(\eta_1,\ldots,\eta_{\bar{S}})$, is represented by $F_\eta(\cdot)$ with support (Supp) $\prod_{i=1}^{\bar{S}} (L_i, U_i)$, where $L_i < \eta_i < U_i$, and U_i and L_i may be unbounded from below and above, respectively. Let $\ell(s)$ be the grade s specific intercept. The X contain no constant. Provided that

(i) $F_\eta(\cdot)$ is absolutely continuous and not a function of X

(ii) $(\eta_1,\ldots,\eta_{\bar{S}})$ is statistically independent of $(X_1,\ldots,X_{\bar{S}})$

(iii) $X_s \in \bar{\bar{X}} \subseteq R^K$ conditional on $X_1 = x_1$, $X_2 = x_2$, $X_{s-1} = x_{s-1}$ for all s, is a K dimensional random variable and does not lie in a proper subspace of R^K (full rank condition)

(iv) Supp $(x_s\beta_s + \ell(s) \mid x_{s-1}\beta_{s-1} + \ell(s\text{-}1),\ldots,x_1\beta_1 + \ell(1)) = (L_s, U_s)$ for $s = 2,\ldots,\bar{S}$, and $\text{Supp}(x_1\beta_1 + \ell(1)) = (L_1, U_1)$.

then F_η is identified and $(\ell(s), \beta_s)$, $s = 1,\ldots,\bar{S}$, are identified up to an affine transformation: $\ell(s) = \text{d}k(s) + \text{e}$, $\beta_s = \text{d}\gamma_s$, where d and e are constants such that $\text{d} > 0$ and e is unrestricted.

Proof: See Appendix A. ■

Assumption (i) guarantees that $F_\eta(\cdot)$ is a proper distribution with a well-defined density and that it is not a function of X_s for any s. Assumption (ii) imposes independence of the shocks from the X_s. The shocks may be freely correlated with each other but not with the X_s. Assumption (iii) rules out linear dependence in X_s and guarantees that X_s has full rank conditional on the preceding values of X_j, and (iv) guarantees that there is some variation in the index $x_s\beta_s + \ell(s)$ across transitions given levels of the preceding indices. With sufficient variation in the X_s, it is possible to identify the entire model up to normalizations. If condition (iii) is weakened so that conditional on $X_{s-1} = x_{s-1}$, $X_{s-2} = x_{s-2},\ldots,X_1 = x_1$, X_s lies in the a proper subspace of R^K, we can only identify linear combinations of the β_s.

Theorem 2 does not cover the important case of no exclusion restrictions (all components of β_s are nonzero, and $X_s = X$, $s = 1,\ldots,m$). This is the standard case considered in the schooling transition literature. Without further restrictions, the model is not identified. However, if β_s, $s = 1,\ldots,m$, where m is the number of continuous variables in X, are a linearly independent set of coefficients (belonging to a subspace of rank m) so that no β_s, can be written as a linear combination of the other values of β_j, some components of the model are identified. In this case, there is an important distinction between components of X that are continuous and those that are discrete. If all components of X are discrete, condition (iv) cannot possibly be satisfied because there are no intervals in the support of $X_s\beta_s + \ell(s)$. The following theorem presents conditions for identification in the case of common X.

Theorem 3. For the schooling model (12a) with $\log c(s \mid x) = \log c(s) - x_s\beta_s + \eta_s$, $s = 1,\ldots,\bar{S}$, assume that distribution of the transition-specific shocks $(\eta_1,..,\eta_{\bar{S}})$, is represented by $F_\eta(\cdot)$ with support (Supp) $\prod_{i=1}^{\bar{S}}(L_i, U_i)$, where $L_i < \eta_i < U_i$, and U_i and L_i may be unbounded. Let $\ell(s)$ be the grade s specific intercept. The X contain no constant. Provided that

(i) $F_\eta(\cdot)$ is absolutely continuous and not a function of X

(ii) $(\eta_1,..,\eta_{\bar{S}})$ is statistically independent of $(X_1,\ldots,X_{\bar{S}})$

(iii) $X_s = X$ for all $S = 1,\ldots,\bar{S}$ is a K dimensional random variable that does not lie in a proper subspace of R^K (full rank condition)

(iv) The first $m \geq 1$ coordinates of X are continuous variables $(m \leq K)$

(v) $\beta_1,\ldots,\beta_m$ are linearly independent

(vi) $\text{Supp}(x_s\beta_s + \ell(s) \mid x_{s-1}\beta_{s-1} + \ell(s\text{-}1),\ldots,x_1\beta_1 + \ell(1)) = (L_s, U_s)$ for $s = 2,..,m$, and $\text{Supp}(x_1\beta_1 + \ell(1)) = (L_1, U_1)$,

then F_η is nonparametrically identified in its first $\text{Min}(m,\bar{S})$ arguments and $(\ell(s),\beta_s)$, $s = 1,\ldots,\text{Min}(m,\bar{S})$

are identified up to an affine transformation $\ell(s) = \mathrm{d}k(s) + \mathrm{e}$, $\beta_s = \mathrm{d}\gamma_s$, where d and e are constants such that d > 0 and e is unrestricted.■

Proof: See Appendix A.

The lack of variation in the X across transitions becomes a very serious constraint on identifiability when the number of continuous variables is small relative to the number of transitions.

When Is the Model Not Identified?

Next consider a case when conditions (v) and (vi) of Theorem 3 are not satisfied and the distribution $F(\eta_1,..,\eta_{\bar{S}})$ is not nonparametrically identified. Variation condition (vi) does not apply to the schooling-transition model estimated in Section 1 when $\beta_s = \beta$ for all s because the same values of the variables appear in all transition equations. Are there alternative conditions that ensure semiparametric identification of this model? Without further restrictions, the answer is "no". We offer the following non-identification result for models that have time invariant $X = x$ when β is fixed across transitions.

Corollary to Theorem 3. Under the conditions of Theorem 3, when conditions (v) and (vi) are not satisfied because $X_1 = X_2 = \ldots = X_{\bar{S}} = X$ and $\beta_1 = \beta_2 = \ldots = \beta_{\bar{S}} = \beta$, but all other conditions of Theorem 2 are valid, $F(\eta_1,\ldots,\eta_{\bar{S}})$ is not identified.

Proof: See Appendix A.■

This corollary implies that identifiability in the schooling-transition model with time-invariant regressors is much more fragile than if there are spell-varying variables. Correcting for "heterogeneity bias" in the schooling transition model framework requires stronger maintained assumptions than are required in an ordered discrete-choice model. In particular, we cannot test the hypothesis of equality of the β across transitions nonparametrically when the regressors are the same in each transition. Because conditions (v) and (vi) of Theorem 3 are violated, we cannot independently vary the arguments of $F(\eta_1,\ldots,\eta_{\bar{S}})$ and trace out the full distribution of the $(\eta_1,\ldots,\eta_{\bar{S}})$.

Sufficient Conditions for Identification

Under the following sufficient conditions on the distribution of the error terms, the schooling-transition model can be identified even when the $X = x$ are identical across transitions. These conditions deliver identification for the logistic schooling-transition model with a discrete factor structure of the type used to produce the estimates reported in Section 1. Heckman and Singer (1984) demonstrate that the nonparametric maximum likelihood estimator of the mixing distribution for η_s or θ is a finite multinomial model with unknown support points of support points. Here we approximate the conditions assumed in constructing their estimator and assume a finite and <u>known</u> upper limit to the number of support points to the number points of support. In practice, if a large upper limit to the number of support points is assumed, there is no difference in their specification of the conditions or ours in finite samples. An alternative mathematically equivalent assumption postulates a known upper limit to the number of types in the population whose choices are invariant across transitions.

Theorem 4. For the model of equation (12c) and (12e) where

$$\log c(s \mid x) = \log c(s) - x\beta_s + \eta_s$$

assume

(i) $\eta_s = \alpha_s \nu + U_s,\ \ s = 1,\ldots,\bar{S}$, where α_s is a transition-specific factor-loading parameter on ν with $\alpha_1 = 1,\ \ |\alpha_s| < \infty\ ,\ s = 1,\ldots,\bar{S}$,

(ii) $(U_1,\ldots,U_{\bar{S}})$ are mutually independent and identically distributed with zero median, known variance, and known log-concave distribution function F (such as normal or logistic)

(iii) ν independent of $(U_1,\ldots,U_{\bar{S}})$

(iv) Supp $(X\beta_s) = \bar{J}_s \subseteq R^1$, an interval in R^1

(v) $X \in \bar{\bar{X}} \subseteq R^K$ does not lie in a proper subspace

(vi) ν is multinomial with an unknown but finite number of mass points ($= I$) (locations and mass probabilities are not known).

Then setting $\alpha_s = 1$ for all s or $\beta_s = \beta$ for all s, or fixing some β_s for a particular s (and letting the associated α_s be free), or fixing some α_s for a particular s (and letting the associated β_s be free), we can identify α_s, β_s, $\ell(s)$, and the distribution of ν (up to a location normalization), where $\ell(s) = c(s)$ plus the location parameter of $\alpha_s \nu$.

Proof: See Appendix A. ■

An unfortunate feature of this theorem is that for each s, only α_s or β_s can be determined. With sufficient variation in X, both can be determined, as shown in the following theorem.

Theorem 5. Define $\bar{I}_s = \text{Supp}(\eta_s)$, the support of η_s. Under the conditions of Theorem 4 with condition (iv) revised to read

(iv)′ $\bar{I}_s \subseteq \text{Supp}(X\beta_s)$ and $X\beta_s$ assumes either arbitrarily large or arbitrarily small values or both and

(vii)′ $\underset{i}{\text{Max}} \, | \nu_i | < \infty$

then α_s, β_s, $\ell(s)$ and the distribution of ν (up to a location normalization) are identified.

Proof: Appendix A. ■

These theorems prove identification of factor structure models for transition models of the type first analyzed in Heckman (1981). Observe that knowledge of F affords identification of β_s, α_s, $s = 1,\ldots,\bar{S}$ if only one component of the X is arbitrarily large even if there are no exclusion restrictions.

Theorems 1-5 establish conditions under which both the ordered-discrete choice model and the schooling transition model are identified. The ordered discrete choice model, which is economically more interpretable than the other models, requires the fewest distributional assumptions. The schooling-transition-model with transition-invariant unobservables assumes myopic decision making and is not nonparametrically identified when the X do not change over transitions. Our analysis indicates that the evidence reported in the empirical literature on educational selectivity rests critically on arbitrary choices of functional forms for distributions of unobservables. Theorems 4 and 5 establish that if a discrete factor structure with a known upper limit for the number of points of support is assumed for the random effect, the schooling-transition model becomes identified by assuming a multinomial distribution for unobservables that persist across spells. This model is consistent with the existence of different types in the population who persistently differ each other in the choices they make over time.

3. Are the Data Consistent with a Structural Schooling Model? Estimates and Implications For Educational Policy

In this section, we compare the empirical performance of the ordered-discrete choice model and the schooling-transition model. We develop implications of our estimated ordered discrete-choice model for educational selectivity and examine the quality of students attracted into college by family income subsidies. Before turning to our estimates of these models, we note that the models of Section 1 were estimated to conform with specifications reported in the published literature. However, we find in both the schooling transition model and the ordered discrete-choice model that a richer empirical specification is indicated. Standard F-tests reported in Table 6 indicate that certain interaction terms noted at the base of the table are statistically important when they are introduced into the models of Section 1.[22]

We then compare three models estimated with these interactions included. Model 1 is the standard logistic schooling-transition model discussed in Section 1.1. Model 2 is the myopic schooling-transition model introduced in Section 1.3 that controls for omitted variables ε and a nonproportional cost function. Essentially, it is the schooling-transition model augmented with person-specific heterogeneity.

The model includes the same variables used in model 1. The transition probability is given by (12e) except that β_s varies by transition. As this model contains the most parameters, it is our benchmark for comparing other models. The distribution of η_s is assumed to satisfy condition (i) of Theorem 3, and the data generating process is assumed to obey the conditions stated in Theorems 4 and 5.

Model 3 is the ordered discrete-choice model of schooling (Section 2.1). It imposes a proportional cost function and assumes no transition-specific shocks affect decision-makers. Individuals observe their endowments of X and ε at birth and choose the level of schooling that maximizes net returns to schooling. This model is simple and easily interpreted. It is consistent with rationality on the part of the agents being studied. It is also the most robust in the sense that distributional assumptions are not needed in order to estimate the model. In this section of the paper, we work with the redefined variable $\omega = -\log \varepsilon$ and the distribution of ω is represented as a two-point mixture of logits for models 2 and 3. This is the two-point version of the nonparametric maximum likelihood estimator presented in Heckman and Singer (1984). The overall density of ability ω is $f(\omega)$ which is

$$f(\omega) = P_1 f_1(\omega) + P_2 f_2(\omega)$$

when $f_1(\omega)$ is the low ability density and $f_2(\omega)$ is the high ability density and P_1 and P_2 are the population proportions of low ability and high ability persons respectively ($P_1 + P_2 = 1$).

Section 3.1 compares the estimated models for five birth cohorts of data. Section 3.2 makes inter-cohort comparisons of structural estimates. Section 3.3 uses the estimated ordered discrete-choice model to examine the effect of income subsidies on enrollment and completion of college and on the quality of students attracted to college by such policies.

3.1 Discriminating Among the Models

To discriminate among competing models, we use goodness-of-fit tests and non-nested model selection tests. Table 7 displays P-values for goodness-of-fit chi-square tests. The test statistics are formed by sorting the data into five equal-size partitions. The predicted probabilities of schooling attainment at each level ("Complete Elementary" through "Attend Graduate School") are compared to actual values for each of the five groups using a chi-square statistic. Panel A displays the P-values when the number of siblings is the sorting criterion, and panel B presents analogous results when the data are sorted by family income. The test criterion is the adequacy of predicting each schooling level for all five groups.[23] Systematic differences in predictions between income groups, due to liquidity constraints for example, should be detected by this procedure. These goodness-of-fit tests do not account for the differences in the number of parameters across the different models and, therefore, tend to favor over-parameterization. Model 1 has 73 more parameters than model 3, and Model 2 has 76 more than model 3.

Comparing row 1 (model 1 - schooling transition model) with row 2 (model 2 - schooling transition model with heterogeneity) in both panels shows that models that control for heterogeneity are favored by the goodness-of-fit tests (higher P-values) although for a few comparisons no clear ranking emerges. Adopting a five-percent significance level, the much more parsimonious ordered discrete-choice model compares favorably against the other two. It is rejected in only one case (see column 2, Panel B). In light of the parsimony of the economic model, it is surprising that it is not rejected more often. It is even more surprising how high the P-values are compared to model 1 P-values when siblings define cells. It is not surprising that P-values for the parsimonious economic model are generally lower than for the unrestricted economic model.[24]

Results from model-selection tests that take into account differences in the numbers of parameters are presented in Table 8. The table presents analogous P-values for test statistics based on the Bayesian Information Criterion (BIC). The distribution of BIC for nonnested models was developed by Vuong (1989).[25] Column 1 and 2 of the table compare the schooling-transition model with and without heterogeneity corrections to the ordered discrete choice model. Results for each cohort are shown in rows 1 to 5. Row 6 shows the difference in the number of parameters. Because the hypotheses being compared are nonnested, the test statistic has three possible outcomes: reject the schooling-transition model in favor of the ordered-discrete-choice model; reject the ordered-discrete choice model in favor of the schooling-transition model; or reject neither. The P-values are for one-sided tests. Negative values indicate that the parsimonious economic model is favored, while positive values indicate that the schooling-transition model is favored. For example, adopting a significance level of 5%, a positive value below .05 means that the parsimonious model is rejected. A negative value between zero and -.05 means that the schooling transition model is rejected, and for other values the models are indistinguishable. Using a .05 significance level, Table 8 shows that ordered-choice model is preferred over the other two models in all cases.

From these analyses we reach the following conclusions. An economically interpretable, stable parameter model can explain data on life-cycle schooling choices. Both the ordered discrete-choice model and the schooling-transition model with heterogeneity controls fit the data in the sense of passing conventional goodness of fit tests. The more highly-parameterized statistical models generally fit the data better but this is not surprising. Using model selection tests that penalize for parameters, the ordered discrete-choice-model is favored over the schooling transition model with or without heterogeneity. The widely-used standard schooling-transition model that does not account for educational selectivity does not fit the data as well as a substantially lower dimensional and more economically interpretable model.

Theorem 1 informs us that the distribution of the unobservables in the ordered discrete choice model can be identified nonparametrically under mild conditions. Theorem 3 and its corollary state that when there are no exclusion restrictions, and the β are assumed constant across transitions, the schooling-transition model with parameters constant across transitions is nonparametrically non-identified. The case of no exclusion restrictions is the situation most commonly encountered in the empirical literature on life cycle schooling attainment. The assumption of a common β across transitions is the benchmark for the educational selectivity hypothesis. Thus the benchmark model is not nonparametrically identified in the commonly encountered case and its empirical content rests on prior assumptions on distributions of unobservables that are imposed on the data. Theorems 4 and 5 inform us that identification in the schooling transition *is* possible when Theorem 2 or Theorem 3 fail to be satisfied if there are a finite number of behavioral types in the population and hence a discrete factor structure model with a known upper limit to the number of discrete mass points is used.[26] However, the ordered-discrete choice model requires fewer *ad hoc* assumptions in order to be estimated, and in this sense is more robust.

The data support the ordered-choice model. It is the simplest model consistent with rational choice that describes the schooling attainment data. The data are <u>not</u> consistent with the educational transition probability models that have been the focal point for analyses in the empirical educational attainment literature.

3.2 Cohort Comparisons and Interpretations of the Ordered Discrete Choice Model

In this subsection we construct the implied ratio of marginal revenue relative to marginal cost for continuing school at each grade level. For an ordered discrete choice model, the marginal lifetime return over the marginal cost of moving from s-1 to s years of schooling is calculated from the estimated threshold value $\ell(s)$:

$$\frac{R(s)-R(s-1)}{c(s)-c(s-1)} = \exp(\ell(s)) \ .$$

Table 9 and Figures 2a and 2b illustrate how the standard deviation of ψ-normalized marginal revenue over marginal cost schedule has changed across cohorts. Implicit in using this diagram to make comparisons across cohorts is the assumption that $\mathrm{Var}(\psi)$ is the same across cohorts. To secure identification of the model, the median and scale of the estimated distribution functions have been set to 0 and 1. Table 9 presents the estimated marginal revenue over marginal cost schedules for each cohort. Standard errors are given in parenthesis. Figure 2a shows graphs of the estimated schedules. They are evaluated at the average family background characteristics of the oldest cohort to isolate changes in the *MR/MC* schedule from shifts in the distribution of family characteristics. By contrast, Figure 2b displays the schedules evaluated at average family background of the corresponding cohort and shows the schedule faced by an "average" person of each generation.

Several patterns are apparent. First, Figure 2a suggests that, except for the most recent cohort, the *MR/MC* schedules have been stable across generations. However, an overall test of global equality of the *MR/MC* parameters for the schedules is rejected at the 5% level (P-values are reported in row 7 of Table 9), though equality of the parameters for the first and third levels of schooling attainment cannot be rejected (row 6). Note that because the scale is unknown, no meaning is to be attached to absolute values of this scale.

Figure 2b shows the *MR/MC* schedules when shifts in the distribution of family characteristics are taken into account. The *MR/MC* schedules display an orderly rise across succeeding generations. Comparing both figures suggests that for all cohorts model but the most recent, it is the shift in the family background characteristics and not a shift in the model parameters that accounts for most of the rise in schooling attainment over the century. The extraordinary rise in the *MR/MC* for the most recent cohort is consistent with the rise in the skill premium documented by Katz and Murphy (1992) and others. Not only did the real wages of lower schooling groups decline, but the real wages of college graduates rose, producing a strong force for raising *MR/MC* and promoting college attendance.

The stability of the estimated schooling demand relationship for the early cohorts flies in the face of conventional wisdom about the effect of the G.I. Bill on boosting post-war college attainment. The GI bill did raise post-war attendance in colleges. However, it did not quite compensate for the reduction in college attendance by males during the second world war. Olson (1974, pp. 46-48) demonstrates that the GI bill nearly restored the deficit between trend-line college enrollment in the absence of World War II and actual enrollment. He estimates the shortfall at 2% over the period 1941-1953. This small discrepancy is not detectable in our data.[27]

3.3 The Effect of Family Income: Short Term Liquidity Constraints or Long Term Factors?

An important implication of the ordered-choice schooling model is that family background and resource variables affect all schooling decisions by altering the ratio of marginal revenue to marginal cost. This section explores the relationship between family income and schooling more closely.

The effects of family income and family size on schooling attainment estimated in the ordered choice model are strong. See the average derivatives of the state probabilities reported in columns 1 and 2 of Table 10 (standard errors are given in parentheses).[28] These derivatives show the effects of a unit increase in a family background variable on the final level of schooling attainment.

An increase in family income, for instance, reallocates persons from lower to higher levels of schooling attainment (column 2). Many people interpret this empirical relationship between schooling and family resources as evidence that *short-term* liquidity constraints affect schooling choices, particularly at the college level. On the basis of such evidence, tuition policies or supplements to family income during the schooling years have been advocated to overcome the alleged failure of credit markets in financing education.

Family income measured in a cross-section represents either short-run resources available to the family or more permanent family influences that may not be easily influenced by cash transfers. Cursory examination of the effects of family size ("Number of Siblings" -- column 1) and "Family Income" (column 2) reveals that their effects on schooling are strong and statistically significant for all cohorts at all levels of schooling.

To determine additional support for which interpretation is more plausible, we use a measure of ability available from the NLSY for the most recent cohort of males. The measure is the AFQT, adjusted for the age at which the test is taken. This measure and its properties have been analyzed in Cawley *et* *al.* (1997). When this variable is entered into the semiparameteric ordered discrete-choice model, the effect of family income on schooling attainment is greatly weakened (compare panels E and F, of Table 10) and is not statistically significantly different from zero in all states. The estimated effects of the other family background measures, especially the effect of father's education are also weakened.[29] This evidence suggests that most of the measured cross-section effect of family income on schooling is due to long-run family background effects that produce AFQT and not liquidity constraints.

There are several possible problems with accepting this interpretation of the evidence without further analysis. First, our measure of family income may contain strong transitory components to which family members do not respond. By a standard errors-in-variables argument, if some average measure of family income is used, introduction of AFQT will not reduce the effect of income if AFQT is a good proxy for permanent income. In a cross-sectional model in which one year of parental income is used, AFQT may proxy permanent family income and nothing more. It may be a better proxy for permanent income than our measure of income. Recall, however, that our measure of family income is a two year average which dampens out some of the transitory components that appear in the more conventional one year measures used in the literature. Second, AFQT may be the result and not the cause of schooling because for some persons in our sample, AFQT is measured after schooling is completed. Conditioning on AFQT may produce spurious empirical results.

The first argument ignores the central point that virtually all of the evidence on the importance of family income on schooling uses a cross-sectional parental income measure. (See for example, Kane, 1994, and the references he cites). We use an average of income over two years, which slightly increases the size and statistical significance of the coefficient on family income compared to the estimated effects when only one year of family income is used.[30] Thus, errors-in-variables may be important, but they do not dominate the parental earnings data. For either the averaged measure or the conventional measure, introduction of AFQT essentially eliminates the influence of parental income on schooling transitions.

The second argument is more serious. To determine if the evidence reported in Panel F of Table 10 is spurious, we use a subset of the data and reestimate the model for a group of persons who were age 14 to 17 and still in school when they took the test. (Most of these students were subject to compulsory schooling laws.) The estimates for this subsample reported in Panel G are, very close to those of Panel E. Conditioning on the predetermined AFQT measure in this subsample (see Panel H) greatly reduces the estimated effect of family income on schooling as before. This is true whether or not a single year family income or two-year average family income is used.

From these analyses, we conclude that the argument that *short-term* credit constraints affect schooling choices is a weak one. Long-term factors like permanent income and family background as crystalized in AFQT play a central role in explaining schooling decisions. Policies that interpret the family income-schooling relationship as arising solely or mainly from liquidity constraints, and hence amenable to short-term tuition or family income subsidies, are misguided. This argument is bolstered by the work of Kane (1994), and Cameron and Heckman (1992, revised 1997), who report evidence that tuition and government subsidies have only small effects on college attendance decisions. Factors more basic than *short-term* cash constraints operating after AFQT is measured determine the schooling-family income relationship.

3.4 Some Consequences of Heterogeneity For Evaluating An Income-Subsidy Policy

Thus far we have not given a name to the unobservable ω whose distribution we estimate in the ordered discrete-choice model. In this subsection, we assume that ω = -log ε is intelligence or aptitude and use the estimates of our model to estimate the effect of an increase of family income on the ability of students attracted into college by such a policy. We examine the effect of a 10% increase in family income on college attendance and graduation rates and estimate where the new students attracted to college come from in the overall ability distribution. Our estimated distribution of ability (ω) is the estimated distribution function of the ordered discrete-choice model. Figure 3 shows how the density changes across the main schooling levels for the most recent cohort, born in 1957-1964. Policies that raise family income attract students from the lower half of the ability distribution into college. These marginal entrants have considerably lower ability than the average student who attends college before family income is increased. Our estimates suggest that analyses of the impacts of educational policies that ignore the heterogeneity in abilities among prospective college students, substantially overstate the gains of such policies for raising the measured incomes of all college graduates. (See Jorgenson and Ho, 1995, for an analysis of this sort). While the students who are induced to enter college will have more income than they would have earned without attending college, they are likely to earn considerably less than the college graduates under the pre-policy regime.

Table 11 presents our evidence on this question. Because only part of ω is properly attributed to ability, our estimates may best be interpreted as upper-bound estimates of the effect of income on changing the composition of ability of students enrolled in college. It must be emphasized, however, that ω is the component of ability that is uncorrelated with observed family background measures. To the extent that ability is a function of family background, our estimates understate the effect of income on the composition of ability among college students because family background is held constant in the simulations below. Panel A presents the parameters of the fitted models for all five cohorts of males.

The first row of Panel A of Table 11 records the proportion of each cohort attending college. The proportion graduating is given in the line labelled 3. Both figures have been rising over time but the growth in attendance is greater than the growth in graduation rates. Using Bayes' theorem, we determine the probability that a person from the upper half of the ability distribution attends college (line 1a) or graduates college (line 3a). For the youngest cohort, 85% of the high ability group attends college but only 47% graduate. Over time, there has been dramatic growth in college attendance among the most able.

The dilution over time in the ability of college students and college graduates is evident from the numbers reported in rows 2a-2b and rows 4a-4b. The average ability of college attenders has declined over cohorts as has the average ability of college graduates. The location of college attenders and college graduates in the overall distribution of ability has also declined over time whether measured by mean ability, (2a for attendees, 4a for graduates), or median ability (2b for attenders, 4b for graduates), although the decline for graduates is not very dramatic.

For the youngest cohorts, most high ability persons attend college. Policies that attract new persons into college mostly attract persons from the lower portion of the ability distribution. To gain some idea of the magnitude of the dilution of quality, consider a substantial 10% increase in family income. Rows 5 and 8 reveal that the response in terms of college enrollment and graduation is modest at best. Row 6 indicates that for the most recent cohort, 69% of the students induced into college come from the low ability distribution, while 21% of the students who graduate come from the low ability distribution (row 9).

The students induced to attend college have dramatically lower ability as measured by their positions in the overall ability distribution (compare rows 2a and 2b with rows 7a and 7b). A similar statement applies to the students induced to graduate college. Using the estimates of the effect of ability on wages for college graduates reported in Cawley *et al.* (1996) and considering only the effect for the youngest cohorts, the wages of the newly induced college graduates (percentile score .71) would be 3-6% lower on average than the mean wages of the original college graduates (percentile score .85). Policies inducing larger responses would show more dramatic differences in mean wages. This evidence suggests that analyses that ignore heterogeneity in student ability may present an overly optimistic account of the likely effect of policies that promote college attendance on the wages of college graduates as a whole. A more precise estimate requires breaking ω into components due to cost and ability. That is a task we leave for another occasion.

4. Summary and Conclusions

This paper examines the economic and econometric foundations of a widely-used statistical model of educational attainment. Studies based on this model have produced an important empirical regularity from many countries with varied economic and social environments: family income and background effects appear to weaken in their effect on coefficients of logit grade transition probabilities at the higher grade levels. We have examined the role of dynamic selection bias or educational selectivity in accounting for this phenomenon.

While the evidence appears to support the educational selectivity or dynamic selection hypothesis when looking at estimated logit coefficients, estimates of average derivatives and average elasticities of transition probabilities do not follow a pattern consistent with the hypothesis. Without an explicit framework to guide interpretation, the evidence is not decisive on this question.

We derive and estimate an interpretable economic model of schooling attainment. It is not consistent with the widely-used schooling transition model. Our framework is consistent with rational decision making by agents. The schooling-transition model implicitly assumes myopic decision making. The ordered discrete-choice model can be estimated invoking fewer parametric restrictions than are required to estimate the schooling transition model and explains the data about as well for five cohorts of American males. We demonstrate that the benchmark schooling transition model with common coefficients across transition is nonparametrically non-identified on the data commonly used in the literature on educational attainment. Distinctions about educational selectivity in the schooling transition model can be made only by invoking intrinsically *ad hoc* distributional assumptions.

Using our model we demonstrate that a substantial portion of the cross-section family income - schooling relationship operating through the return/cost function is due to permanent family environmental factors including levels of permanent income and the host of factors that produce measured AFQT and is not due to short-run credit constraints. Interpreting the unobservable in the ordered discrete-choice model as ability, we find that policies designed to induce college attendance and college graduation by increasing family income attract substantially less able persons to attend and graduate college. Failure to account for such heterogeneity in human ability leads to an overly optimistic assessment of the effect of supplements to family income in raising the aggregate stock of skills.

References

Amemiya, Takeshi. Advanced Econometrics. Cambridge, Mass.: Harvard University Press, 1985.

Bartholemew, D. J. Stochastic Models For Social Processes. 2nd Edition, Wiley, 1973.

Blossfeld, H. and Shavit, Y. Persistent Inequality: Changing Educational Attainment in Thirteen Countries, Boulder, Co: Westview Press, 1993.

Bretagnolle, J. and C. Huber-Carol. "Effects of Omitting Covariates in Cox's Model For Survival Data." Scandinavian Journal of Statistics 15 (Winter 1988): 125-138.

Cameron, S. and Heckman, J. "The Dynamics of Educational Attainment Among Blacks, Hispanics, and Whites." Manuscript. Chicago: University of Chicago, Department of Economics, 1992, revised 1997.

_____ "Life-Cycle Schooling and Educational Selectivity: Models and Evidence, a Complete Version." Manuscript. Chicago: University of Chicago, Department of Economics, 1994.

Cameron, S. and C. Taber. "Evaluation and Identification of Semiparametric Maximum Likelihood Models of Dynamic Discrete Choice." Manuscript. Chicago: University of Chicago, Department of Economics, February, 1994.

Cawley, J., J. Heckman, L. Lochner and E. Vytlacil. "Measuring The Effects of Cognitive Ability." Manuscript. Chicago: University of Chicago, Department of Economics, April, 1995 revised May, 1996.

Comay, Y., A. Melnik, and M. Pollatschek. "The Option Value of Education and the Optimal Path for Investment in Human Capital." International Economics Review 14 (June 1973): 421-35.

Cosslett, S. "Distribution-Free Maximum Likelihood Estimator For The Binary Choice Model." Econometrica 51 (May 1983): 765-782.

Featherman, R. and R. M. Hauser. "Design For A Replicate Study of Social Mobility in the United States." In Social Indicator Models, edited by Kenneth C. Land and Seymour Spilerman, New

York: Russell Sage Foundation, 1975.

Follman, D. "Nonparametric Mixtures of Logistic Regression Models." Ph.D. dissertation, Carnegie Mellon, 1985.

Follman, D. and Diane Lambert. "Generalizing Logistic Regression By Nonparametric Mixing." Journal of American Statiatical Association 48 (March 1989): 295-300.

Duncan, O. D. "A Socioeconomic Index For All Occupations." In Occupations and Social Status. edited by Albert J. Reiss, Jr., Glencoe, N.Y.: Free Press, 1961.

Hansen, W. L. "Impact of Student Financial Aid on Access." The Crisis in Higher Education, edited by J. Froomkin, New York: Academy of Political Science, 1983.

Heckman, J. "Statistical Models For Discrete Panel Data." In Structural Analysis of Discrete Data, edited by C. Manski and D. McFadden, Cambridge: MIT Press, 1981.

_____, "The Chi-Square Goodness of Fit Statistics for Models with Parameters Estimated from Microdata." Econometrica 52 (November 1984): 1543-47.

Heckman, J. and Borjas, G. "Does Unemployment Cause Future Unemployment? Definitions, Questions and Answer from a Continuous Time Model of Heterogeneity and State Dependence." Economica 47 (August 1980): 247-283.

Heckman, J. and S. Burton. "A Method For Minimizing The Impact of Distributional Assumptions." Econometrica 52 (March 1984): 271-320.

Heckman, J. and Honoré, B. "The Empirical Content Of The Roy Model." Econometrica 85 (September 1990): 1121-49.

Heckman, James, Lance Lochner and Petra Todd. "Choice-Theoretic Generalizations of the Ordered Discrete Choice Model." Manuscript, Chicago: University of Chicago, June, 1996.

Ichimura, H. "Semiparametric Least Squares (SLS) and Weighted SLS Estimation of Single-Index Models." Journal of Econometrics 58 (July 1993): 71-120.

Jorgenson D. and Min Ho. "Policies To Raise The Stock of Human Capital." Kuznets Lectures, Yale University. Manuscript, Harvard University, 1995.

Kane, T. J. "College Entry by Blacks since 1970: The Role of College Costs, Family Background and The Returns to Education." Journal of Political Economy 102 (October 1994): pp. 878-911.

Karlin, S. Total Positivity, Stanford, California: Stanford University Press, 1968.

Katz, L. and K. Murphy. "Changes in Relative Wages, 1963-1987: Supply and Demand Factors." Quarterly Journal of Economics 101 (February 1992): 35-78.

Klein, R. and Spady, R. "An Efficient Semiparametric Estimator for Binary Response Models." Econometrica 61 (March 1993): 387-421.

Matzkin, R. "Semiparametric Estimation of Monotone and Concave Utility Functions For Polychotomous Choice Models." Econometrica 59(5) (September 1991): 1315-1327.

Manski, C. "Identification of Binary Reponse Models". Journal of American Statistical Association 83 (August 1988): 729-737.

Manski, C. and Wise, D. College Choice in America. Cambridge, Mass: Harvard University Press, 1983.

Mare, R. D., "Social Background and School Continuation Decisions." Journal of The American Statistical Association 75: (June, 1980): 295-305.

_____, "Change and Stability in Educational Stratification." American Sociological Review 46 (January 1981): 72-87.

Olson, K. The G.I. Bill, The Veterans and the Colleges. Lexington: University of Kentucky Press, 1974.

Ridder, G. "The Non-Parametric Identification of Generalized Accelerated Failure-Time Models." Review of Economic Studies 57 (April 1990): 167-81.

Sewell, W. and Hauser, R. Education, Occupation, and Earnings: Achievement in the Early

Career. New York: Academic Press, 1975.

Spilerman, S. "Careers, Labor Market Structure, and Socioeconomic Achievement." American Journal of Sociology 83 (November 1977): 551-93.

Vuong, Q. "Likelihood Ratio Tests for Model Selection and Non-nested Hypotheses." Econometrica 57, (March 1989): 307-333.

Weisbrod, B. "Education and Investment in Human Capital." Journal of Political Economy Supplement (October 1962): 106-123.

Willis, R. and Rosen, S. "Education and Self-Selection." Journal of Political Economy 87 Part 2 (October 1979): S7-S36.

Yakowitz, S. J. and Spragins, J. D. "On The Identifiability of Finite Mixtures." Annals of Mathematical Statistics (January 1968): 209-214.

This research was supported by NSF 87-10145, NSF-91-11455, NSF-SBR-93-0248 and grants from the Mellon and Spencer Foundations. An early version of this paper was presented at an Institute For Research on Poverty Workshop in June, 1990. We are grateful to Shubham Chaudhuri, Bo Honoré, Michael Keane, Tom MaCurdy, Bob Mason, Rosa Matzkin, Derek Neal, Jose Scheinkman, Michael Rothschild, Joe Tracy, Ed Vytlacil and two anonymous and very helpful referees, for comments. We also thank Chris Taber for comments and programming assistance.

Endnotes

1. A semi-Markov model of the sort discussed by Bartholomew (1973) or Spilerman (1977) can easily be adapted to accomodate time in the state (duration dependence) as an explanatory variable in the grade transition framework. Time-varying or age-varying explanatory variables may therefore be introduced as well. But as noted in Heckman and Borjas (1980), *ad hoc* remedies for introducing time-varying explanatory variables can induce severe statistical problems. One remedy to these problems is to reformulate the state space as a grade-attained-by-age (or time) state space. See Cameron and Heckman (1992, revised 1997) for a model that does this.

2. The data are drawn from the Occupational Changes in a Generation Survey (OCG), collected as a special supplement to the March 1973 Current Population Survey (further details are reported in Appendix B). Table 1 reports estimates for White American males born in 1937-1946.

3. For similar reasons, a fully nonparametric and unrestricted simultaneous equations system cannot be identified.

4. For example, homogeneity of degree one, as in Matzkin (1992).

5. If Θ is discrete (*e.g.* binomial), sums replace integrals and probabilities of the discrete values of Θ replace $F(\theta)$ in expression (5).

6. For discussion and proof of (7), see Appendix A. The expression is an average value of the score vector of Θ projected on the score vector of X.

7. Mare (1980) presents an incorrect analysis of the consequences of misspecification in this model. His error is that he assumes that classical specification error analysis for linear regression analysis applied to the logit $\log(P/1\text{-}P) = X\beta$ characterizes misspecification of a logit model fit on microdata. The Mare analysis applies when P lies strictly between 0 and 1 but not when P is a binary random variable. Equation (7) in the text provides the right characterization of specification bias when Θ is omitted and the model is fit using maximum likelihood on microdata. See Cameron and Heckman (1994) for further details.

8. The result on the variances is a consequence of the log-concavity of the densities of X and Θ used in this example. See Heckman and Honoré (1990).

9. The vanishing bias for the binomial case follows from the fact that the support of the distribution is bounded and not from the fact that Θ is a discrete random variable.

10. In particular, we try to ensure comparability with the widely-cited work of Mare (1980). Our analysis differs from Mare's earlier work only because we delete the "Duncan Index" from the list of regressors used by Mare. The "Duncan Index", though in widespread use in sociology, is a measure of "occupational prestige" obtained from surveys of individual attitudes toward different occupations and is difficult to interpret apart from the highly-correlated elements income and education. An appendix available on request from the authors presents estimates from models that include the Duncan Index. None of our main statistical inferences are affected by its exclusion. In section 3, we present evidence that a model with interactions of income and education is more consistent with the evidence.

11. This procedure relies on multinomial distributions with an unknown number of points to approximate the unknown distribution of the omitted variable Θ. An alternative interpretation of their estimator is as a multinomial model for unobservables. See Theorems 4 and 5 below. Subject to identification and

regularity conditions, this procedure produces consistent parameter estimates. For all OCG and NLSY cohorts, multinomial distributions with two or three points of support were determined to be adequate approximations to the unknown distribution of Θ.

12. Much of the previous literature has focused on comparisons of estimated β_s from the same transitions across cohorts or estimates of different transitions within a cohort. Logit slope parameters are viewed as "pure inequality of opportunity" measures for schooling transitions and the size of β_s is seen to measure the dependence of transition probabilities s on family background variables. See Mare (1981), for example.

13. These results are not included here but can be found in an appendix to Cameron and Heckman (1994), and are available on request from the authors.

14. The term ε can represent omitted variable as did the Θ heterogeneity term of Section 1, if we let $\varphi(x) = \exp(x\beta) = \exp(x_o\beta_o + x_u\beta_u) = \varphi(x_o)\varepsilon$, where $\varepsilon = \exp(x_u\beta_u)$ and "o" and "u" represent the observed and unobserved parts of the X vector.

15. The weak convexity imposed in defining this model is not an intrinsic feature of the schooling choice problem and is relaxed in Heckman, Lochner and Todd (1996).

16. Strictly speaking, it is not necessary to make a specific functional form assumption for φ to identify or estimate the model although we do so in this paper. Matzkin (1992) demonstrates that φ can be nonparametrically identified up to scale using conventional postulates for cost functions.

17. If $\log\varepsilon$ were known to have a normal distribution with variance $\sigma^2_{\log\varepsilon}$,

$$Pr(S=s|X=x) = \Phi\left[\frac{\ell(s-1) + x\beta}{\sigma_{\log\varepsilon}}\right] - \Phi\left[\frac{\ell(s) + x\beta}{\sigma_{\log\varepsilon}}\right],$$

where Φ is the univariate standard normal distribution. Hence, an ordered-probit model could be used to recover estimates of parameters $\ell(s)$ and β.

18. This theorem is an extension of a result in Manski (1988) for binary discrete-choice problems.

19. This condition can be weakened to allow for quantile independence, but we do not do so in this paper.

20. The exact nature of the option value depends on the specific assumptions of the model. Comay, Melnik, and Pollatschek (1973) discuss option values of schooling building on the work of Weisbrod (1962).

21. In the perfect foresight case, (12a) could be justified by assuming that the υ_j follow a "no news is good news" process, where

$$\upsilon_j \leq \upsilon_{j+1} \leq \upsilon_{j+2} \quad j = 0,\ldots,\bar{S}\text{-}2.$$

However, this assumption does not justify an independent error term. As the X variables do not vary by grade, errors in data measurement also cannot justify the error term as the errors would then be identical across transitions and could not be statistically independent.

22. Estimates of the models of Section 1 based on the full set of interactions do not reverse any conclusions reported there. These estimates are available on request from the authors.

23.The chi-square statistics are adjusted for the variance of the estimated parameters in the predicted probabilities (Heckman, 1984).

24. We used a mixing-distribution approach to approximate the true $F(\cdot)$. This approach is similar to the one used to estimate the heterogeneity distributions in the schooling-transition model. In addition, we experimented with a kernel-based estimator similar to the ones for discrete choice proposed by Ichimura (1993) and Klein and Spady (1993). The kernel estimator provided a slightly better fit to the data but was computationally burdensome, so our reported estimates were all calculated using the simpler mixing-distribution estimator.

25. Vuong establishes the distribution of this criterion assuming that both models are misspecified.

26. This approach originates in Heckman (1981).

27. We thank Richard Campbell for the Olson reference.

28. Similarly strong estimates are found in the other schooling models. See Cameron and Heckman (1994) for these results.

29. This argument is developed more fully in Cameron and Heckman (1992, revised 1997) who use other methods to account for schooling-AFQT simultaneity bias.

30. These results are available on request from the authors.

Table 1

Educational Transition Probabilities for OCG White Males Born 1937-1946 (Age 26 - 35 in 1973)
Logistic Probability Estimates (t-values)*

	(1) Complete Elementary	(2) Attend High School	(3) Graduate High School	(4) Attend College	(5) Graduate College	(6) Attend 17+
Number of Siblings	-0.086(3.35)	-0.142(4.82)	-0.147(9.58)	-0.110(7.52)	-0.081(3.62)	0.031(0.95)
Family Income at Age 16	0.198(7.78)	0.149(5.32)	0.071(6.50)	0.069(9.46)	0.029(3.72)	-0.009(0.93)
HGC Father	0.131(4.28)	0.110(3.32)	0.089(5.45)	0.085(6.85)	0.016(1.04)	0.009(0.47)
HGC Mother	0.190(5.84)	0.100(2.85)	0.116(6.54)	0.124(8.62)	0.101(5.39)	0.087(3.37)
Broken Home	-0.118(0.49)	-0.051(0.19)	-0.260(2.04)	-0.064(0.57)	-0.178(1.18)	-0.503(2.32)
Farm Residence at Age 16	-0.092(0.48)	-0.133(0.61)	0.392(2.99)	-0.181(1.76)	-0.039(0.27)	-0.080(0.39)
Southern Birth	-0.109(0.61)	0.293(1.50)	-0.003(0.03)	0.016 (0.20)	-0.146(1.51)	-0.494(3.73)
Intercept	0.247(0.84)	1.394(3.89)	-0.050(0263)	-2.226(13.41)	-1.252(5.76)	0.771(2.52)

*See the base of Table 4 for variable definitions. A data description can be found in Appendix B.

Table 2

How the Distribution of X and θ_0, Across Grade Transitions

A. Logit Model with $\theta \sim N(0,1)$

Grade Transition	(1) E(x)	(2) E(θ)	(3) Corr(x,θ)	(4) Var(x)	(5) Var(θ)
1	0.00	0.00	.00	1.00	1.00
2	0.36	0.36	-.15	0.87	0.87
3	0.57	0.57	-.25	0.80	0.80
4	0.71	0.71	-.30	0.77	0.77
5	0.82	0.82	-.35	0.74	0.74
10	1.11	1.39	-.45	0.69	0.69
50	1.75	1.75	-.63	0.63	0.63
B. Logit Model with Binomial θ, E(θ) = 0, Var(θ) = 1					
1	0.00	0.00	.00	1.00	1.00
2	0.36	0.39	-.16	0.87	0.85
3	0.55	0.61	-.25	0.80	0.63
4	0.69	0.73	-.28	0.75	0.47
5	0.80	0.80	-.30	0.71	0.36
10	1.17	0.93	-.26	0.57	0.14
50	2.20	1.00	-.09	0.35	0.01

Note: These moments are computed from the joint distribution of θ and X conditional on survival. See equation (8) for the conditional density. Initially, X is distributed N(0,1).

Table 3

Estimated β_s for Ten Grade Logistic Transition Model using Simulated Data* (True β_s = 1)

A. $\theta \sim N(0, 1)$

Transition	(1) $\hat{\beta}_s$	(2) Monte Carlo Std. Dev. of $\hat{\beta}_s$
1	0.84	0.4
2	0.73	0.6
3	0.64	0.7
4	0.60	0.8
5	0.56	.10
10	0.47	.14
B. $\theta \sim$ Binomial, $E(\theta) = 0$, $Var(\theta) = 1$		
1	0.82	0.3
2	0.68	0.4
3	0.66	0.6
4	0.69	0.9
5	0.72	0.9
10	0.80	.13

*Data Generation: The model that generated the data was specified as $U_s = \beta_0 + X\beta_s + \theta + \varepsilon_s$ where $d_s = 1$ if $U_s \geq 0$ and $d_s = 0$ otherwise. If $d_s = 0$, then a "death" occurs and schooling ends at that transition. The X's were drawn from a N(0,1) distribution. Each ε_s was drawn independently from a logistic distribution. In the binomial case, the heterogeneity term, θ, was drawn from a binomial distribution with equal mass at -1 and 1, so the overall variance of the heterogeneity was 1. The slopes and intercepts in each transition were set to 1 and 0 respectively to force the restriction that $E(\theta) = 0$. Each β_s was also assumed to be 1.

For estimation, the slopes and intercepts were not constrained to be equal across transitions, so separate estimates of the intercept and β_s are obtained at each transition. Only estimates of the slope are reported above. One-hundred runs were estimated, each with 5000 initial cases. Standard maximum likelihood techniques (logits) were used to estimate the model.

Table 4

Educational Transition Probabilities for OCG White Males Born 1937-1946 (Age 26 to 35 in 1973)
Logistic Probability Estimates (t-values) with a Nonparametric Heterogeneity Correction

	(1) Complete Elementary	(2) Attend High School	(3) Graduate High School	(4) Attend College	(5) Graduate College	(6) Attend 17+
(1) Number of Siblings	-0.101 (3.8)	-0.159 (5.1)	-0.175 (10.0)	-0.163 (8.5)	-0.175 (5.6)	-0.036 (0.8)
(2) Family Income at Age 16	0.201 (7.7)	0.156 (5.5)	0.075 (6.6)	0.082 (9.6)	0.054 (5.2)	0.024 (1.6)
(3) HGC Father	0.145 (4.6)	0.130 (3.8)	0.110 (6.3)	0.120 (7.7)	0.064 (3.1)	0.069 (2.4)
(4) HGC Mother	0.201 (5.9)	0.117 (3.2)	0.144 (7.3)	0.175 (9.2)	0.190 (7.0)	0.207 (5.5)
(5) Broken Home at Age 16	-0.124 (0.5)	-0.037 (0.1)	-0.304 (2.2)	-0.134 (1.0)	-0.309 (1.5)	-0.786 (2.7)
(6) Farm Residence at Age 16	-0.039 (0.2)	-0.089 (0.4)	0.432 (3.0)	-0.209 (1.6)	-0.109 (0.6)	-0.147 (0.5)
(7) Southern Birth	-0.055 (0.3)	0.368 (1.8)	0.038 (0.4)	0.012 (0.1)	-0.235 (1.9)	-0.762 (4.1)

Notes: *Family Income is denominated in 1000's of 1995 dollars. A two-point model was deemed sufficient to characterize the heterogeneity distribution.

Variable Definitions: "Family Income Age 16" is the income of the individual's parents in the individual's 16th year: "HGC Father" and "HGC Mother" are the highest grade attained by the individual's father and mother: "Broken Home": is a binary variable indicating whether one or more of the individual's parents were absent from his household most of the time up to age 16; "Farm Residence" is an indicator recording whether the individual lived on a farm at age 16; "Southern Birth" records whether or not the individual was born in the Southern Census region.

Table 5

OCG White Males Born 1937-1946 (Age 26 to 35 in 1973)

A. Average Derivatives of the Transition Probabilities with Respect to the Explanatory Variables
(Derived from Parameter Estimates presented in Table 1)

Transition	(1) Number of Siblings	(2) Family Income	(3) HGC Father	(4) HGC Mother	(5) Broken Home	(6) Farm Residence	(7) Southern Birth
(1) Complete 8	-0.002	0.005	0.003	0.005	-0.004	-0.005	-0.003
(2) Attend HS	-0.003	0.004	0.003	0.002	-0.001	-0.003	0.007
(3) Complete HS	-0.014	0.007	0.009	0.011	-0.025	0.038	-0.000
(4) Attend College	-0.023	0.014	0.017	0.025	-0.013	-0.037	0.003
(5) Graduate College	-0.019	0.007	0.004	0.024	-0.042	-0.009	-0.034
(6) Attend 17+	0.008	-0.002	0.002	0.021	-0.121	-0.019	-0.119

B. Average Total Elasticities. Partial Effects in Parentheses (Parameter Estimates in Table 1)

Transition	(1) Number of Siblings	(2) Family Income	(3) HGC Father	(4) HGC Mother	(5) Broken Home	(6) Farm Residence	(7) Southern Birth
(1) Complete 8	-.015 (-.015)	.010 (.010)	.006 (.006)	.011 (.011)	-.000 (-.000)	-.001 (-.001)	-.001 (-.001)
(2) Attend HS	-.032 (-.017)	.015 (.005)	.010 (.004)	.015 (.004)	-.001 (-.000)	-.002 (-.001)	.000 (.002)
(3) Complete HS	-.167 (-.135)	.058 (.044)	.074 (.064)	.019 (.094)	-.007 (-.007)	.011 (.014)	.000 (-.000)
(4) Attend College	-.607 (-.440)	.422 (.363)	.606 (.532)	.967 (.858)	-.014 (-.007)	-.025 (-.037)	.005 (.005)
(5) Gradute College	-.916 (-.309)	.619 (.198)	.729 (.123)	1.798 (.831)	-.036 (-.022)	-.032 (-.007)	-.042 (-.048)
(6) Attend 17+	-.831 (.085)	.556 (-.063)	.797 (.067)	2.459 (.661)	-.097 (-.061)	-.043 (-.011)	-.198 (-.155)

C. Heterogeneity-Corrected Average Total Elasticities of the Transition Probabilities. Partial Effects in Parentheses
(Parameter Estimates in Table 4)

Transition	(1) Number of Siblings	(2) Family Income	(3) HGC Father	(4) HGC Mother	(5) Broken Home	(6) Farm Residence	(7) Southern Birth
(1) Complete 8	-.018 (-.018)	.012 (.012)	.010 (.010)	.015 (.015)	-.000 (-.000)	-.000 (-.000)	-.001 (-.001)
(2) Attend HS	-.034 (-.019)	.018 (.007)	.016 (.007)	.022 (.007)	-.000 (-.000)	-.001 (-.001)	.002 (.002)
(3) Complete HS	-.142 (-.128)	.058 (.049)	.095 (.086)	.142 (.128)	-.007 (-.007)	.011 (.011)	.003 (.002)
(4) Attend College	-.415 (-.383)	.321 (.306)	.541 (.516)	.857 (.821)	-.011 (-.009)	-.023 (-.025)	.003 (.002)
(5) Graduate College	-.745 (-.461)	.400 (.237)	.567 (.303)	1.375 (.953)	-.030 (-.024)	-.031 (-.013)	-.050 (-.051)
(6) Attend 17+	-.743 (-.087)	.342 (.103)	.658 (.316)	1.851 (.998)	-.106 (-.082)	-.047 (-.020)	-.254 (-.214)

TABLE 6

P-Values of Likelihood Ratio Tests
for the Significance of Interaction Terms*

Model/Cohort	(1) 1958-65	(2) 1938-47	(3) 1928-37	(4) 1918-27	(5) 1908-1917
(1) Ordered-Choice Model	.000	.000	.000	.000	.000
(2) Schooling Transition Model (no heterogeneity correction)	.000	.000	.000	.000	.021
(3) Schooling Transition Model (with heterogeneity correction)	.000	.000	.000	.000	.012

NOTE: There are eight degrees of freedom for the tests of ordered-choice models and forty-eight for the schooling-transition model tests (that is, the eight interactions enter into 6 transition equations).

*Eight interactions and squared terms of the original variables were added to the specification reported in Section 1: number of siblings squared, family income squared, a dummy variable indicating whether highest grade completed of the father is ≥ 13, a dummy variable indicating that the highest grade completed of the mother is ≥ 13, family income interacted with born home, family income interacted with farm residence, highest grade completed of the father interacted with farm residence, and highest grade completed of father interacted with Southern birth. This set of higher-order interactions was chosen from a larger set comprised of all non-redundant interaction terms plus indicator variables for father's education greater than high school and mother's education greater than high school. Variables not significant in at least three of the five cohorts were deleted from the set until the eight interaction terms listed above remained.

Table 7

P-Values of Goodness-of-Fit Tests (Chi-Square Tests)

Birth Cohort	(1) 1957-64 (NLSY)	(2) 1937-46 (OCG)	(3) 1927-36 (OCG)	(4) 1917-26 (OCG)	(5) 1907-16 (OCG)
A. 5 Cells Sorted by Number of Siblings					
Model 1. Schooling Transition Model without Heterogeneity Correction	.94	.32	.09	.54	.04
Model 2. Schooling Transition Model with Heterogeneity Correction	.98	.75	.82	.68	.13
Model 3. Ordered-Discrete-Choice Model	.96	.76	.19	.16	.70
B. 5 Cells Sorted by Family Income					
Model 1. Schooling Transition Model without Heterogeneity Correction	.97	.59	.09	.81	.03
Model 2. Schooling Transition Model with Heterogeneity Correction	.98	.60	.19	.78	.41
Model 3. Ordered-Discrete-Choice Model	.76	.21	.01	.07	.16

TABLE 8

Non-nested Tests of Model Selection
P-Values of a Standard-Normal Distribution
Test Statistic Based on Bayesian Information Criterion*
(all models include interaction terms)

Birth Cohort (Data)	(1) Transition Model (No Heterogeneity) vs. Ordered-Choice Model	(2) Transition Model with Heterogeneity vs. Ordered-Choice Model
1. 1908-17 (OCG)	-.00	-.00
2. 1918-27 (OCG)	-.02	-.04
3. 1928-37 (OCG)	-.00	-.00
4. 1938-47 (OCG)	-.00	-.00
5. 1957-65 (NLSY)	-.00	-.00
6. Difference in Number of Parameters	73	76

*P-values are based on Vuong's model selection test for non-nested models (Vuong, 1989, Theorem 5.1) with a correction for the number of parameters in each model. The test statistic, $[(p-q)\ \ell n\ n]/[2\sqrt{n}\ w_n$ where p and q are the number of parameters in each model, n is the sample size and w_n is the estimated standard error of the log-likelihood ratio (see Vuong, 1989, p. 318).

Notes: This non-nested test has three possible outcomes: both models are equal, reject the ordered-choice model in favor of the schooling-transition model, and reject the schooling-transition model in favor of the ordered-choice model. p-values presented are for one-sided tests, and negative values indicate that the ordered-choice model is favored, while positive values indicate that the schooling-transition model is favored. For example, adopting a significance level of 10%, a positive value below .10 means that the ordered-choice model is rejected in favor of the transition model; a negative value between zero and -.10 means the schooling-transition model is rejected, and other values indicate that both models are statistically indistinguishable at the 10% level. The -.00 in row 1 of Column 1 indicates that the transition model is rejected in favor of the ordered-choice model.

Table 9

Estimated *MR/MC* Schedules for Each Birth Cohort Weighted by the Average Characteristics of the Oldest Cohort (Compare Figure 2a)
(standard errors in parentheses)

Birth Cohort	Complete 8	Attend HS	Grad HS	Attend College	Grad Coll	Attend 17+
(1) 1907-1916	2.963	1.634	0.917	0.391	0.221	0.123
	(.233)	(.124)	(.070)	(.033)	(.021)	(.015)
(2) 1917-1926	2.528	1.559	0.834	0.299	0.164	0.086
	(.259)	(.157)	(.088)	(.037)	(.022)	(.013)
(3) 1927-1936	2.650	1.835	0.969	0.360	0.200	0.111
	(.287)	(.188)	(.097)	(.037)	(.022)	(.013)
(4) 1937-1946	2.725	1.931	0.862	0.279	0.131	0.077
	(.350)	(.236)	(.104)	(.034)	(.017)	(.010)
(5) 1957-1964	3.555	2.807	1.177	0.433	0.170	0.091
	(.837)	(.661)	(.285)	(.113)	(.050)	(.030)
P-Values of Chi-Square Statistics for Equality of the *MR/MC* Schedules at Each Schooling State*						
(6)	.19	.01	.21	.01	.00	.00
P-Value of a Chi-Square Statistic for Overall Equality of the *MR/MC* Schedules*						
(7)	.01					

*Notes: The chi-square statistics for the state-by-state tests are constructed by forming a vector of all pairwise differences in the estimated parameters and calculating the associated covariance matrix. The chi-square statistic is then formed in the usual way. With five cohorts there are 10 pairwise differences. Since only 4 are linearly independent, the test statistic is distributed as a $\chi^2(4)$. The test for overall equality is constructed analogously except that pairwise differences of a single parameter are replaced by pairwise differences in the vector of parameters representing the *MR/MC* schedule for each of the 6 schooling states.

Table 10
Average Derivatives of State Probabilities* (Standard Errors)

State	Numb Sibs	Family Income	HGC Father	HGC Mother	Broken Home	Farm Res	Southern Birth	AFQT Score
				A. 1908-1917 Birth Cohort (OCG Males)				
Less than 8	.0130	-.0034	-.0117	-.0115	.0450	.0881	.0277	-
	(.0011)	(.0003)	(.0012)	(.0013)	(.0098)	(.0080)	(.0079)	-
Complete 8 Only	.0086	-.0021	-.0067	-.0069	.0225	.0383	.0040	-
	(.0008)	(.0002)	(.0019)	(.0013)	(.0065)	(.0041)	(.0041)	-
Attend HS Only	.0050	-.0023	-.0036	-.0047	.0074	.0081	-.0082	-
	(.0008)	(.0001)	(.0015)	(.0021)	(.0048)	(.0037)	(.0031)	-
Complete HS Only	-.0061	0.0026	.0013	.0044	-.0249	-.0493	-.0211	-
	(.0008)	(.0002)	(.0009)	(.0024)	(.0051)	(.0061)	(.0051)	-
Attend Coll Only	-.0078	.0023	.0030	.0053	-.0211	-.0370	-.0067	-
	(.0007)	(.0001)	(.0007)	(.0010)	(.0055)	(.0040)	(.0037)	-
Grad Coll Only	-.0061	.0015	.0043	.0050	-.0143	-.0239	.0001	-
	(.0007)	(.0001)	(.0011)	(.0015)	(.0048)	(.0033)	(.0029)	-
Attend 17+	-.0067	.0013	.0079	.0072	-.0146	-.0242	.0042	-
	(.0010)	(.0001)	(.0027)	(.0023)	(.0061)	(.0056)	(.0035)	-
				B. 1918-1927 Birth Cohort (OCG Males)				
Less than 8	.0079	-.0015	-.0084	-.0074	.0219	.0519	.0279	-
	(.0007)	(.0002)	(.0008)	(.0008)	(.0059)	(.0052)	(.0048)	-
Complete 8 Only	0.0067	-.0013	-.0045	-.0056	.0159	.0291	.0133	-
	(.0005)	(.0001)	(.0005)	(.0008)	(.0038)	(.0031)	(.0028)	-
Attend HS Only	.0100	-.0012	-.0048	-.0077	.0203	.0285	.0091	-
	(.0008)	(.0001)	(.0010)	(.0015)	(.0046)	(.0044)	(.0037)	-
Complete HS Only	.0018	.0007	-.0018	-.0032	-.0014	-.0208	-.0190	-
	(.0012)	(.0002)	(.0010)	(.0022)	(.0035)	(.0067)	(.0044)	-
Attend Coll Only	-.0076	.0012	.0034	.0053	-.0170	-.0319	-.0151	-
	(.0008)	(.0001)	(.0008)	(.0010)	(.0042)	(.0039)	(.0032)	-
Grad Coll Only	-.0081	.0011	.0058	.0071	-.0174	-.0271	-.0096	-
	(.0008)	(.0001)	(.0010)	(.0014)	(.0040)	(.0038)	(.0031)	-
Attend 17+	-.0106	.0011	.0103	.0109	-.0222	-.0297	-.0066	-
	(.0016)	(.0001)	(.0030)	(.0031)	(.0058)	(.0069)	(.0043)	-
				C. 1928-1937 Birth Cohort (OCG Males)				
Less than 8	.0069	-.0014	-.0069	-.0054	-.0004	.0378	.0170	-
	(.0006)	(.0002)	(.0007)	(.0008)	(.0052)	(.0047)	(.0039)	-
Complete 8 Only	.0048	-.0011	-.0035	-.0035	.0011	.0156	.0048	-
	(.0004)	(.0001)	(.0003)	(.0007)	(.0030)	(.0024)	(.0018)	-
Attend HS Only	.0119	-.0021	-.0055	-.0071	.0043	.0239	.0023	-
	(.0010)	(.0002)	(.0010)	(.0023)	(.0064)	(.0055)	(.0039)	-
Complete HS Only	.0084	-.0002	-.0060	-.0058	.0065	-.0096	-.0178	-
	(.0016)	(.0002)	(.0018)	(.0023)	(.0049)	(.0068)	(.0044)	-
Attend Coll Only	-.0079	.0020	-.0002	.0041	-.0011	-.0255	-.0097	-
	(.0008)	(.0002)	(.0008)	(.0012)	(.0045)	(.0041)	(.0030)	-
Grad Coll Only	-.0092	.0014	.0056	.0055	-.0031	-.0193	-.0021	-
	(.0009)	(.0001)	(.0013)	(.0022)	(.0048)	(.0042)	(.0029)	-
Attend 17+	-.0149	.0017	.0161	.0124	-.0073	-.0229	.0055	-
	(.0018)	(.0002)	(.0033)	(.0044)	(.0083)	(.0085)	(.0051)	-

(continued)

	D. 1938-1947 Birth Cohort (OCG Males)							
Less than 8	.0044	-.0012	-.0035	-.0050	.0057	.0007	.0022	-
	(.0004)	(.0001)	(.0004)	(.0006)	(.0028)	(.0029)	(.0022)	-
Complete 8 Only	.0031	-.0008	-.0019	-.0045	.0046	-.0009	.0011	-
	(.0003)	(.0001)	(.0003)	(.0005)	(.0018)	(.0017)	(.0011)	-
Attend HS Only	.0123	-.0024	-.0087	-.0181	.0199	-.0077	.0028	-
	(.0011)	(.0002)	(.0014)	(.0028)	(.0067)	(.0063)	(.0038)	-
Complete HS Only	.0121	-.0011	-.0118	-.0136	.0242	-.0149	-.0005	-
	(.0019)	(.0002)	(.0040)	(.0057)	(.0078)	(.0095)	(.0053)	-
Attend Coll Only	-.0079	.0019	.0057	.0045	-.0104	.0023	-.0028	-
	(.0011)	(.0002)	(.0016)	(.0025)	(.0045)	(.0042)	(.0029)	-
Grad Coll Only	-.0078	.0014	.0064	.0093	-.0133	.0056	-.0014	-
	(.0008)	(.0001)	(.0018)	(.0020)	(.0043)	(.0042)	(.0024)	-
Attend 17+	-.0162	.0021	.0160	.0253	-.0307	.0149	-.0014	-
	(.0021)	(.0002)	(.0040)	(.0060)	(.0099)	(.0108)	(.0056)	-
	E. 1958-1965 Birth Cohort (NLSY Males)							
Less than 8	.0040	-.0003	-.0034	-.0046	.0038	-.0096	.0121	-
	(.0007)	(.0001)	(.0015)	(.0014)	(.0035)	(.0054)	(.0029)	-
Complete 8 Only	.0019	-.0002	-.0016	-.0027	.0012	-.0029	.0045	-
	(.0003)	(.0001)	(.0008)	(.0007)	(.0016)	(.0020)	(.0010)	-
Attend HS Only	.0133	-.0007	-.0128	-.0162	.0050	-.0092	.0232	-
	(.0021)	(.0002)	(.0064)	(.0056)	(.0110)	(.0121)	(.0063)	-
Complete HS Only	.0185	-.0009	-.0194	-.0319	-.0016	.0109	.0129	-
	(.0030)	(.0002)	(.0084)	(.0075)	(.0151)	(.0185)	(.0093)	-
Attend Coll Only	-.0072	.0004	-.0046	-.0047	-.0085	0.0223	-.0254	-
	(.0030)	(.0002)	(.0054)	(.0047)	(.0065)	(.0123)	(.0067)	-
Grad Coll Only	-.0110	.0006	.0113	.0178	-.0022	.0030	-.0151	-
	(.0019)	(.0002)	(.0047)	(.0055)	(.0092)	(.0101)	(.0054)	-
Attend 17+	-.0195	.0011	.0283	.0399	.0024	-.0146	-.0121	-
	(.0044)	(.0003)	(.0115)	(.0107)	(.0166)	(.0213)	(.0099)	-
	F. 1958-1965 Birth Cohort wiht AFQT Score (NLSY Males)							
Less than 8	.0027	.0000	-.0013	-.0020	.0051	-.0014	.0087	-.0015
	(.0006)	(.0001)	(.0011)	(.0008)	(.0032)	(.0058)	(.0026)	(.0002)
Complete 8 Only	.0011	.0000	-.0007	-.0010	.0017	.0001	.0028	-.0015
	(.0002)	(.0001)	(.0006)	(.0004)	(.0013)	(.0019)	(.0009)	(.0001)
Attend HS Only	.0074	-.0001	-.0036	-.0073	.0078	.0043	.0118	-.0006
	(.0016)	(.0006)	(.0043)	(.0034)	(.0086)	(.0098)	(.0051)	(.0001)
Complete HS Only	.0112	-.0003	-.0113	-.0175	.0047	.0150	.0039	-.0006
	(.0025)	(.0002)	(.0082)	(.0055)	(.0136)	(.0171)	(.0084)	(.0001)
Attend Coll Only	-.0019	-.0002	-.0127	-.0089	-.0077	.0071	-.0148	-.0006
	(.0031)	(.0002)	(.0047)	(.0036)	(.0048)	(.0122)	(.0056)	(.0001)
Grad Coll Only	-.0066	.0001	.0037	.0083	-.0058	-.0051	-.0084	-.0006
	(.0018)	(.0002)	(.0037)	(.0025)	(.0081)	(.0092)	(.0048)	(.0001)
Attend 17+	-.0140	.0004	.0256	.0284	-.0057	-.0199	-.0038	-.0006
	(.0045)	(.0002)	(.0120)	(.0106)	(.0172)	(.0220)	(.0105)	(.0001)

(continued)

	G. 1963-1965 Birth Cohort (NLSY Males Less than Age 17 in 1979)							
Less than 8	.0042	-.0003	-.0044	-.0051	.0079	-.0200	.0115	-
	(.0011)	(.0001)	(.0016)	(.0015)	(.0053)	(.0085)	(.0041)	-
Complete 8 Only	.0019	-.0003	-.0023	-.0024	.0034	-.0066	.0038	-
	(.0006)	(.0001)	(.0008)	(.0011)	(.0024)	(.0033)	(.0017)	-
Attend HS Only	.0144	-.0011	-.0160	-.0157	.0245	-.0331	.0194	-
	(.0029)	(.0004)	(.0072)	(.0063)	(.0171)	(.0172)	(.0089)	-
Complete HS Only	.0168	-.0011	-.0250	-.0257	.0269	-.0134	.0062	-
	(.0041)	(.0003)	(.0107)	(.0103)	(.0213)	(.0247)	(.0106)	-
Attend Coll Only	-.0066	.0005	-.0013	-.0015	-.0114	.0345	-.0217	-
	(.0022)	(.0003)	(.0013)	(.0043)	(.0078)	(.0176)	(.0082)	-
Grad Coll Only	-.0117	.0009	.0100	.0142	-.0193	.0224	-.0125	-
	(.0025)	(.0002)	(.0046)	(.0053)	(.0140)	(.0138)	(.0066)	-
Attend 17+	-.0191	.0015	.0245	.0343	-.0319	.0161	-.0066	-
	(.0044)	(.0003)	(.0081)	(.0120)	(.0250)	(.0309)	(.0126)	-
	H. 1963-1965 Birth Cohort with AFQT (NLSY Males Less than Age 17 in 1979)							
Less than 8	.0029	-.0001	-.0023	-.0037	.0084	-.0088	.0071	-.0016
	(.0009)	(.0002)	(.0010)	(.0010)	(.0050)	(.0079)	(.0037)	(.0003)
Complete 8 Only	.0012	-.0000	-.0015	-.0022	.0032	-.0022	.0019	-.0006
	(.0003)	(.0001)	(.0006)	(.0007)	(.0020)	(.0024)	(.0012)	(.0001)
Attend HS Only	.0077	-.0000	-.0053	-.0133	.0200	-.0081	.0068	-.0037
	(.0024)	(.0002)	(.0020)	(.0052)	(.0136)	(.0130)	(.0069)	(.0004)
Complete HS Only	.0097	-.0004	-.0110	-.0270	.0234	.0009	-.0033	-.0042
	(.0032)	(.0003)	(.0059)	(.0102)	(.0183)	(.0208)	(.0099)	(.0004)
Attend Coll Only	-.0032	-.0003	-.0070	-.0062	-.0085	.0116	-.0123	.0018
	(.0037)	(.0003)	(.0046)	(.0055)	(.0112)	(.0149)	(.0069)	(.0014)
Grad Coll Only	-.0066	.0001	.0043	.0136	-.0166	.0050	-.0037	.0031
	(.0023)	(.0002)	(.0023)	(.0059)	(.0123)	(.0116)	(.0063)	(.0005)
Attend 17+	-.0117	.0006	.0218	.0384	-.0298	.0016	.0035	.0053
	(.0049)	(.0003)	(.0094)	(.0132)	(.0242)	(.0260)	(.0117)	(.0015)

*Note.—Estimated From the ordered discrete-choice model. Family income is denominated in 1000's of 1995 dollars. All models include the interaction terms listed at the base of Table 6.

Table 11

Comparisons of Average and Marginal College Entrants and Graduates

A. Baseline Statistics*

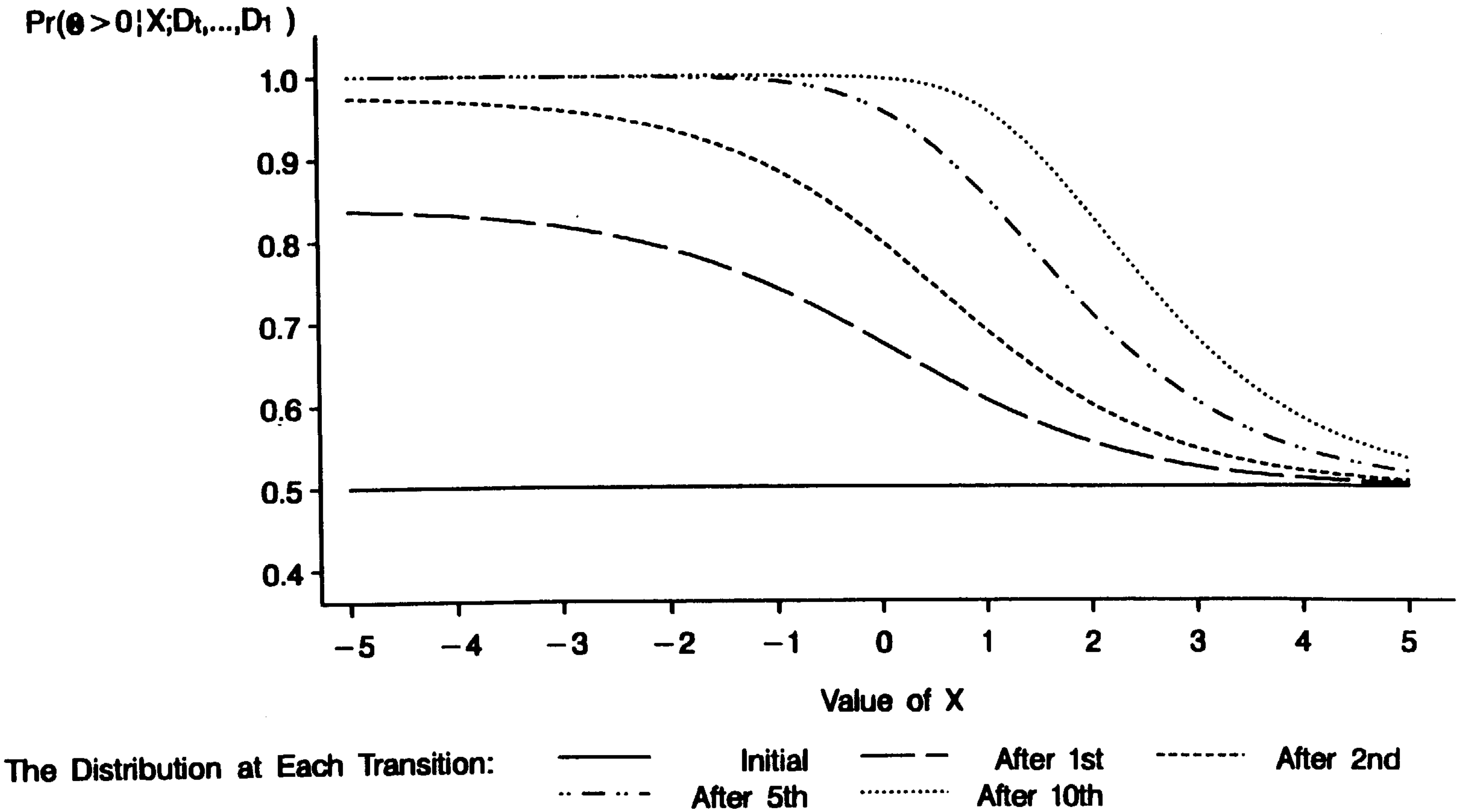
Birth Cohort (Data)	(1) 1957-1963 (NLSY)	(2) 1937-1946 (OCG)	(3) 1927-1936 (OCG)	(4) 1917-1926 (OCG)	(5) 1907-1916 (OCG)
(1) Probabillity of attending college	.56	.47	.40	.32	.24
(1a) Probability of high-ability* person attending college	.85	.76	.69	.58	.44
(2a) Location of average ability of college attenders in the baseline distributional (percentile)	.73	.76	.80	.83	.88
(2b) Location of median ability of college attenders in the baseline distribution (percentile)	.71	.75	.78	.81	.85
(3) Probability of graduating college	.26	.25	.23	.17	.13
(3a) Probability of high-ability* person graduating college	.47	.46	.43	.33	.25
(4a) Location of average ability of graduates in the baseline distribution (percentile)	.85	.86	.89	.90	.92
(4b) Location of median ability of graduates in the baseline distributional (percentile)	.85	.86	.87	.89	.91
B. Increase Family Income 10%					
(5) Rise in probability of college attendance	.008	.008	.011	.007	.007
(6) Fraction of increased enrollment from lower-half of ability distribution	.69	.48	.42	.28	.22
(7a) Location of average ability of new entrants in the baseline distribution (percentile)	.52	.56	.63	.64	.71
(7b) Location of median ability of new entrants in the baseline distribution (percentile)	.50	.55	.62	.62	.70
(8) Rise in probability of college graduation	.008	.011	.008	.004	.004
(9) Fraction of increased graduation from lower half of ability distribution	.21	.15	.18	.15	.16
(10a) Location of average ability of new graduates in the baseline distribution	.71	.73	.72	.79	.84
(10b) Location of median ability of new graduates in the baseline distribution	.70	.72	.71	.78	.83

Table continues on next page

Notes: The median and variance of all the baseline (unconditional) distributions are normalized to 0 and 1 respectively. "High Ability" is defined as an ω = $-\log\varepsilon$ greater than 0 (that is, it is in the upper half of the distribution). Row (1) corresponds to $\Pr(S \geq s)$, where "s" denotes a level of schooling and "S" represents a random variable indicating schooling attainment. Row (1a) is $\Pr(S \geq s \mid \omega > 0)$. Row (2a) reports the percentile of the baseline distribution $F(\omega)$ (unconditional) that corresponds to the value of $E(\omega \mid S \geq s)$, and row (2b) displays the percentile of the baseline distribution that corresponds to the value of $\text{Median}(\omega \mid S \geq s)$. Rows (3)-(4b) present the analogous statistics for college graduation.

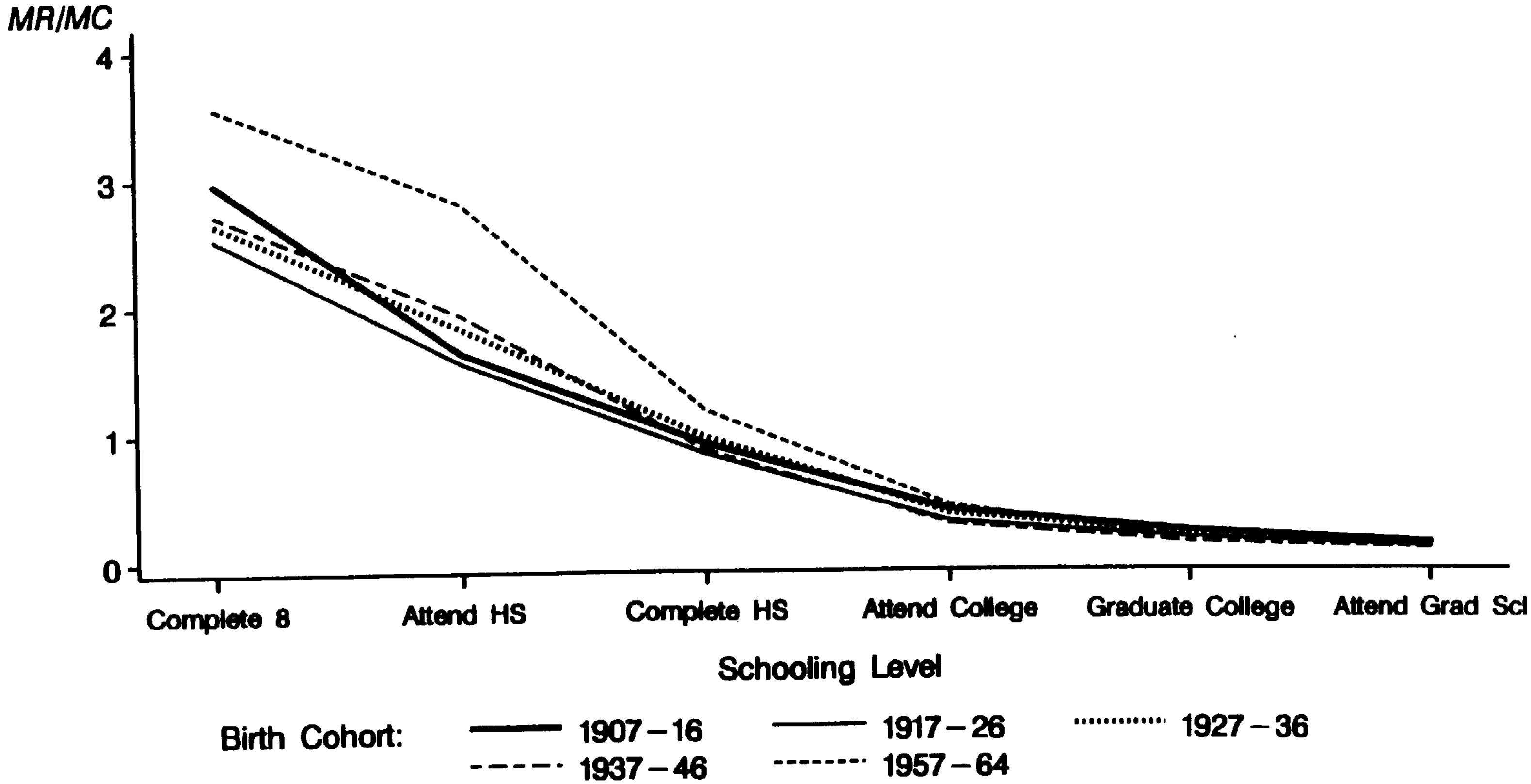
Row (5) of panel A shows the rise in the probability of college attendance due to an increase in family income. Row (6) displays the fraction of (5) that comes from the population with $\omega < 0$. Rows (7a) and (7b) report the location of average ability and median ability of new entrants in the baseline distribution. Rows (8) - (10b) display analogous figures for college graduation.

Figure 1. The Probability of $\theta > 0$, $Pr(\theta > 0 | X; D_t = 1, \ldots, D_1 = 1)$, for a Logistic Schooling–Transition Model with N(0,1) Heterogeneity – Plotted against X.



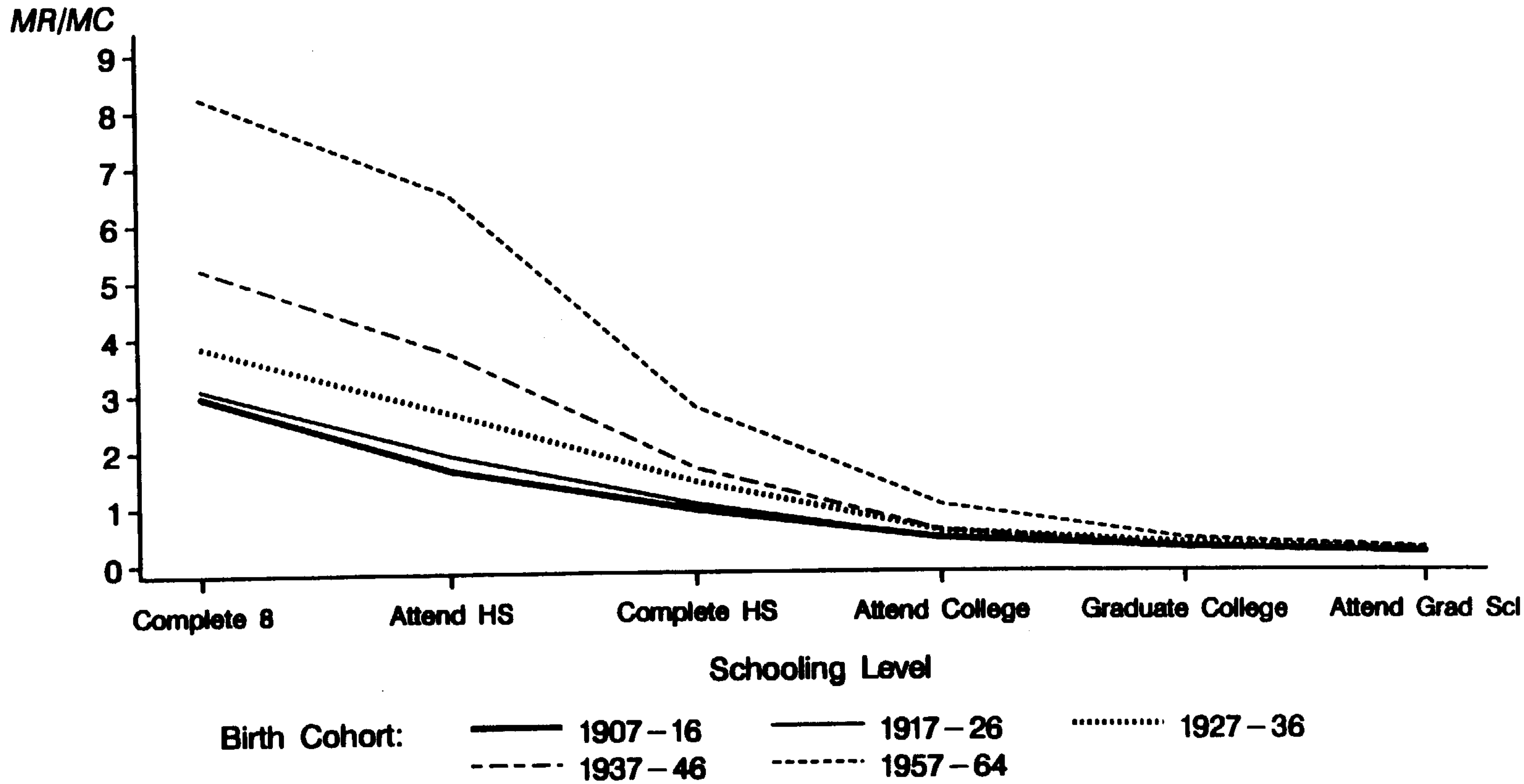
Note: X and θ, the heterogeneity term, are initially independent. See the text for more details.

Figure 2a. The Marginal Revenue/Marginal Cost Schedule for an Increment in Schooling. Each schedule is evaluated at the oldest cohort's average family characteristics. (Constructed from estimates of the ordered discrete–choice model)



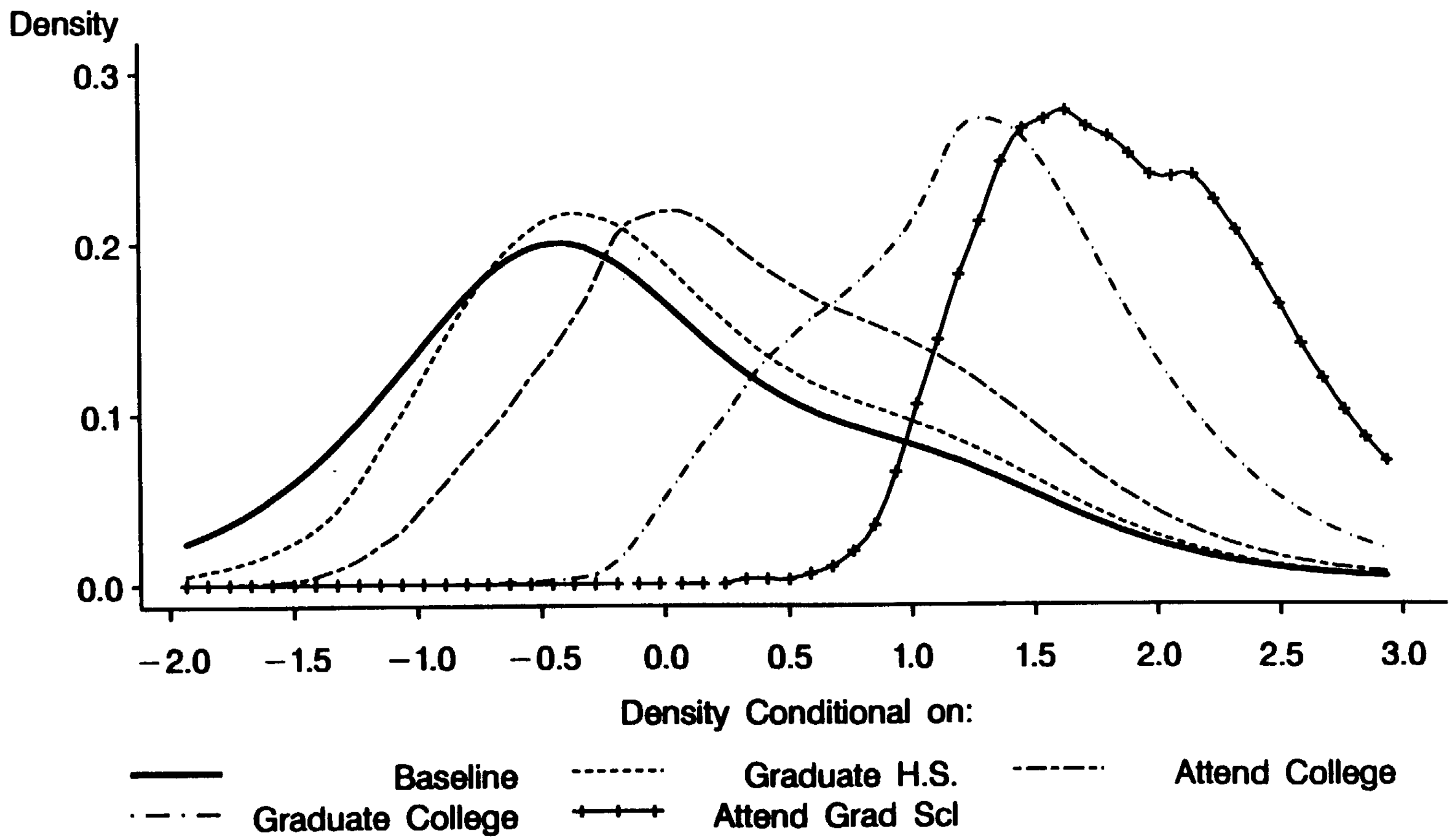
Note: See equations (9) and (10) for a precise statement of "Marginal Revenue over Marginal Cost." The median and scale of the underlying distribution functions are normalized to 0 and 1.

Figure 2b. The Marginal Revenue/Marginal Cost Schedule for an Increment in Schooling. Each schedule is evaluated at average family characteristics of the cohort. (Constructed from estimates of the ordered discrete–choice model)



Note: See equations (9) and (10) for a precise statement of "Marginal Revenue over Marginal Cost." The median and scale of the underlying distribution functions are normalized to 0 and 1.

Figure 3. Density of ω (Ability) for Ordered-Choice Model at Various Schooling Levels. For Whites Males Born 1957-1964 (NLSY).



Note: The median and scale of the baseline distribution are normalized to 0 and 1 respectively.

Appendix A: Proofs

Proof of Equation (7)

Let F be a log concave cdf. This includes the logit, probit, and linear probability models as special cases. Let $Pr(D = 1 \mid X = x) = F(x\beta)$. Partition X into (X_o, X_u), observed and unobserved components, with associated coefficients (β_o, β_u). Let K be the number of observed components, and let $X_u\beta_u = \Theta$. Then we may rewrite the model as

$$Pr(D = 1 \mid X = x) = F(x_o\,\beta_o + \theta).$$

The "true" coefficient associated with θ is "1". Let the coefficient associated with θ be denoted by τ. To simplify the notation delete the "o" subscript on β and X. Assume random sampling and standard regularity conditions as in Amemiya (1985), Theorem (4.1.1) so that if N is the sample size and if θ is observed, then

$$\pounds^{(N)} = \frac{1}{N}\sum_{i=1}^{N} [D_i \log F(x_i\beta + \tau\theta)] + (1 - D_i) \log(1 - F(x_i\beta + \tau\theta))]$$

$$\underset{a.s}{\rightarrow} \pounds(\beta,\tau) \ ,$$

which is concave in (β, τ) and assumed to be at least twice continuously differentiable.

Following a line of argument whose essential idea is due to Bretagnolle and Huber (1988), consider the following constrained optimization problem. Maximize $\pounds^{(N)}$ for various values of $\tau \in [0,1]$. The value $\tau = 0$ corresponds to omission of θ from the model. For every choice of τ, define $\beta_O(\tau)$. From concavity of $\pounds(\beta, \tau)$, $\beta_O(\tau)$ is well-defined and unique. With probability one, the sample estimate $\hat{\beta}_O(\tau)$ converges to $\beta_O(\tau)$.

The population information matrix is

$$- \nabla^2 \pounds(\beta,\tau) = \begin{bmatrix} I_{OO} & I_{OU} \\ I_{UO} & I_{UU} \end{bmatrix}$$

where

$$I_{00} = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} E_{X_i\theta} \left[\frac{-\partial^2 \pounds_i^{(N)}}{\partial\beta_0 \, \partial\beta_0'} \right] , \quad I_{UU} = \lim_{N\to 0} \frac{1}{N} \sum_{i=1}^{N} E_{X_i\theta} \left[\frac{-\partial^2 \pounds_i^{(N)}}{\partial\tau \, \partial\tau'} \right] ,$$

$$I_{0U} = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} E_{X_i\theta} \left[\frac{-\partial^2 \pounds_i^{(N)}}{\partial\beta_0 \, \partial\tau'} \right] = I_{OU}' \text{ , and}$$

$\pounds_i^{(N)}$ is the contribution to $\pounds^{(N)}$ from the ith observation. The limits are assumed to exist.

Let $I_{00}(\tau)$, $I_{0U}(\tau)$ denote matrices evaluated at $(\beta_0(\tau),\tau)$. By concavity, $\beta_0(\tau)$ is the unique solution to

$$\nabla \pounds \, (\beta_0(\tau),\tau) = 0.$$

Differentiating the first order conditions with respect to τ, we obtain

$$\frac{d\beta}{d\tau}(\tau) = -\left[I_{00}(\tau)\right]^{-1} \left[I_{0U}(\tau)\right] .$$

Integrating over the interval [0,1], we obtain

$$\beta(1) - \beta(0) = -\int_0^1 \left[I_{00}(\tau)\right]^{-1} I_{0U}(\tau) d\tau$$

which is expression (7) with $\beta(1) = \beta_1$ and $\beta(0) = \gamma_1$ using the logit cdf for F.

Analogous expressions can be obtained for any misspecified higher-order transition equation if log-concavity of the F holds. Simply evaluate the information matrix using the appropriate grade-specific joint density of (Θ, X) as derived in the text. ■

Proofs of Theorems 1 to 4

Proof of Theorem 1.

By hypothesis, $Pr(D_j = 1 \mid X = x)$ $j = 1,\ldots,\bar{S}$ are known for all values of X. Further, (A-1) $$Pr(D_1 = 1 \mid X = x) = F_\psi(\ell(1) + x\beta)$$

(A-2) $$Pr(D_j = 1 \mid X = x) = F_\psi(\ell(j) + x\beta) - F_\psi(\ell(j\text{-}1) + x\beta), \quad j = 2,\ldots,\bar{S} - 1$$

(A-3) $$Pr(D_{\bar{S}} = 1 \mid X = x) = 1 - F_\psi(\ell(\bar{S} - 1) + x\beta).$$

Following the analysis of Manski (1988, Proposition 2, Corollary 5), we may treat (A-1) as a binary discrete choice model. Under conditions (i), (ii) and (iii) in Theorem 1, he proves that if there is no constant among the X we may identify $\ell(1)$, β, and F_ψ up to scale $d > 0$, provided that we make one additional quantile assumption (*e.g.* Median $\psi = 0$ for quantile $\alpha = ½$). With such an assumption, define a set of values

$$\bar{\bar{X}}_\alpha = \{X \mid Pr(D_1 = 1 \mid X = x) = 1 - \alpha\}.$$

We then use the quantile assumption to solve for

$$\ell(1) = Q_\alpha(\psi) - X_\alpha\beta$$

up to scale d where $X_\alpha \in \bar{\bar{X}}_\alpha$ and $Q_\alpha(\psi)$ is the known quantile. Without the quantile assumption, $\ell(1)$ is not identified, even up to scale.

Using (A-2), we may write

$$Pr(D_2 = 1 \mid X = x) = F_\psi(\ell(2) + x\beta) - F_\psi(\ell(1) + x\beta).$$

Rearranging,

$$Pr(D_2 = 1 \mid X = x) + F_\psi(\ell(1) + x\beta) = F_\psi(\ell(2) + x\beta)$$

where the terms on the left-hand side are known. Since F_ψ and β are known (up to scale d), we can determine $\ell(2)$ up to scale from $\ell(1)$ and β. Assuming monotonicity of F over some interval (ψ_ℓ, ψ_u),

$$\ell(2) = F_\psi^{-1}[Pr(D = 1 \mid X = x) + F_\psi(\ell(1) + x\beta)] - x\beta.$$

By similar reasoning, and proceeding sequentially through equations A(2), we can determine $\ell(j)$, $j = 1,\ldots,\bar{S}$ up to affine transformations.■

Remark:

In the semiparametric ordered discrete-choice model with more than two intervals, we gain an important identifying feature not present in the binary discrete-choice model. Let $g(x;\beta)$ be the deterministic portion of the index function. Under additive separability a person experiences outcome i if

(A-4) $$\ell(i\text{-}1) \leq g(x;\theta) + \psi < \ell(i), \qquad i = 1,\ldots,I.$$

Only positive affine transformations of g, $\ell(i)$, ψ preserve the probability content of (A-4). Thus let $T{:}R^1 \rightarrow R^1$ be a monotonic and increasing transformation. The probability of the events in (A-4) is preserved if T is applied to ψ and $\ell(i) - g(x;\beta)$; that is,

$$Pr(\ell(i\text{-}1) - g(x;\beta) \leq \psi < \ell(i) - g(x;\beta)) = Pr(T(\ell(i\text{-}1) - g(x;\beta)) \leq T(\psi) < T(\ell(i) - g(x;\beta))).$$

However, for the monotonic transformation to apply to $\ell(i)$ and $g(x;\beta)$ separately and still preserve the probability content, it is required that

$$T(\ell(i) - g(x;\beta)) = T(\ell(i)) - T(g(x(\beta)),$$

that is

$$T(x + y) = T(x) + T(y)$$

so that T is an affine transformation. Arbitrary monotonic transformations that are admissible in the binary discrete-choice problem are not admissible in this context because they do not preserve the inequality in (A-4) when they are applied to each component of the inequality. (Compare the more general class of admissible transformations given in Manski, Appendix A). Ridder (1990) presents an alternative proof of Theorem 1.

Proof of Theorem 2.

By hypothesis, we know the left hand sides of the following $\bar{S}$ equations:

(A-5-1) $Pr(D_1 = 1 \mid X_1 = x_1) = F_{\eta_1}(\ell(1) + x_1\beta_1)$

(A-5-2) $Pr(D_1 = 1, D_2 = 1 \mid X_1 = x_1, X_2 = x_2) = F_{\eta_1,\eta_2}(\ell(1) + x_1\beta_1, \ell(2) + x_2\beta_2)$

...

(A-5-3) $Pr(D_1=1, D_2=1,\ldots,D_{\bar{S}} = 1 \mid X_1=x_1,\ldots,X_{\bar{S}} = x_{\bar{S}}) = F_{\eta_1,\eta_2,\ldots,\eta_{\bar{S}}}(\ell(1) + x_1\beta_1,\ldots,\ell(\bar{S})+x_{\bar{S}}\beta_{\bar{S}})$.

Following the first step in the proof of Theorem 1, we can use (A-5-1) to identify F_{η_1}, $\ell(1)$, β_1 up to scale (where we can normalize by Var(η_1) or by some other constant such as $\|\beta_1\|$). A condition like Median (η_1) = 0 secures identification of $\ell(1)$ up to scale (recall that X_1 is assumed to contain no constant). Note that we use (iii) to identify β_1 up to a scalar transformation.

The proof proceeds recursively. Using (A-5-2), let $(\ell(1) + x_1\beta_1) \to U_1$. This produces $F_{\eta_2}(\ell(2)+x_2\beta_2)$ for values defined in the limit set. From assumption (iii), the limit set imposes no restriction on the support of x_2 and we may repeat the preceding argument and claim that F_{η_2}, $\ell(2)$ and β_2 are identified up to a scale convention (*e.g.* Var (η_2) or $\|\beta_2\|$ set equal to an arbitrary constant). The location of $\ell(2)$ is identified using an assumption about a quantile or first moment of η_2. Our ability to vary $\ell(2) + x_2\beta_2$ and freely to trace out F_{η_2} follows from assumption (iv). Proceeding in this fashion, we can recover $\ell(j) + x_j\beta_j$ $j = 1,\ldots,\bar{S}$ subject to scale and location normalizations. Recovery of the joint distribution of $(\eta_1,\ldots,\eta_{\bar{S}})$ follows from knowledge of the $1 \times \bar{S}$ vector $G = (\ell(1)+x_1\beta_1,\ldots,\ell(\bar{S})+x_{\bar{S}}\beta_{\bar{S}})$ where $G_j = \ell(j) + x_j\beta_j$. By varying the components of G, we can trace out the joint distribution $F_{\eta_1,\ldots,\eta_{\bar{S}}}$. ■

Proof of Theorem 3.

Without further restrictions on β, condition (iv) of Theorem 2 cannot be satisfied if $X_s = X$. The special case of $\beta_s = \beta$, where F is not identified, is treated in the corollary to this Theorem which is proved below.

To identify parameters beyond those obtained from step 1 of Theorem 2 observe that the left hand side of

$$Pr(D_1 = 1, D_2 = 1 \mid X = x) = F_{\eta_1,\eta_2}(\ell(1) + x\beta_1, \ell(2) + x\beta_2)$$

is known as is the first argument on the right hand side of the expression. Let

$$\ell(1) + x\beta_1 = G_1$$

and recall that the first m coordinates of x correspond to the continuous regressors. Without loss of generality, assume $m \geq 1$, and $\beta_{11} \neq 0$. Then we can write

$$x_1 = \frac{G_1}{\beta_{11}} - x_2\frac{\beta_{12}}{\beta_{11}} - \dots - x_K\frac{\beta_{1K}}{\beta_{11}} - \frac{\ell(1)}{\beta_{11}}$$

where in this expression lower case x_i is the i^{th} coordinate of x.

In the index $x\beta_2$, use standard Gaussian elimination and substitute for x_i, from the preceding equation and obtain

$$\left[\frac{G_1}{\beta_{11}} - x_2\frac{\beta_{12}}{\beta_{11}} - \dots - x_K\frac{\beta_{1K}}{\beta_{11}} - \frac{\ell(1)}{\beta_{11}}\right]\beta_{21} + \beta_{22}x_2 + \dots + \beta_{2K}x_K + \ell(2).$$

These variables can be freely varied given $x\beta_1 = G_1$. Proceeding recursively, in the $(j+1)$th argument, $(j < m)$, we obtain an expression that substitutes out for $(x_1,\dots,x_j)$ leaving m-j free continuous variables.

Array the β_j into a matrix B with the j^{th} row of B being β_j. B is an $\bar{S} \times K$ matrix. Let $B(r,n)$ be the $r \times n$ submatrix of B consisting of the first r rows and n columns, and let $B(r, K - n)$ be the matrix consisting of the first r rows and the last K-n columns of B. Partition β_j into the first e elements $(\beta_j(e))$

and the last K-e elements $\beta_j(K - e)$.

In this notation, successive Gaussian elimination produces

$$\tilde{\beta}_{j+1} = \beta_j(K-j) - \beta_{j+1}(j)[B(j,j)]^{-1}B(j,K-j)$$

a K-j dimensional vector. In order for $[B(j,j)]^{-1}$ to exist, it is necessary that $\beta_1,\ldots,\beta_j$ be linearly independent vectors. Condition (v) assures us that this requirement is satisfied for $j \leq m$. Define $\widetilde{\beta}_{j+1}(m-j)$ as the first $(m$-$j)$ elements of $\widetilde{\beta}_{j+1}$ associated with the continuous regressors. In order to satisfy (vi), at least one component of $\widetilde{\beta}_{j+1}(m\text{-}j)$ must be non-zero, $j \leq \text{Min}[m,\bar{S}] - 1$. Define $\hat{x}^j = (x_j,\ldots,x_K)$.

Let G_1 vary so that $x_1\beta_1 + \ell(1) = G_1 \rightarrow U_1$. We obtain the marginal distribution

$$F_{\eta_2}(\ell_2^* + \hat{x}^2\tilde{\beta}_2)$$

where $\ell_2^* = (\ell(2) - \dfrac{\ell(1)}{\beta_{11}}\beta_{21})$ is obtained via the same linear transformation that is used to obtain $\widetilde{\beta}_2$.

Invoking (vi), we continue in this fashion and identify $F_{\eta_1},\ldots,F_{\eta_m}$,as well as β_1, $\widetilde{\beta}_2,\ldots,\widetilde{\beta}_m$ and $\ell(1)$, $\ell_2^*,\ldots,\ell_m^*$, up to scale and location using the same reasoning as was used in the proof of Theorem 2 applied to the transformed variables. We may normalize the $\widetilde{\beta}_2$ to obtain the scale and use location assumptions (*e.g.*, Median $\eta_i = 0$) to define the intercepts. Define

$$G_1 = \ell(1) + x\beta_1,$$

$$G_j = \ell_j^* + \hat{x}^j\ \widetilde{\beta}_j,\ j = 2,\ldots,m$$

and using

$$Pr(D_1=1,\ldots,D_m=1\,|\,X=x) = F_{\eta_1,\ldots,\eta_m}(\ell(1) + x\beta_1, \ell_2^* + \hat{x}^2\tilde{\beta}_2,\ldots,\ell_m^* + \hat{x}^j\tilde{\beta}_m)$$

find, for each element in x, the associated G_i values and the associated probability. Proceeding in this fashion, we trace out the full distribution of $F(\eta_1,..,\eta_m)$.

Using the fact that

$$G_j = \ell_j^* + \hat{x}^j \tilde{\beta}_j = \ell(j)+x\beta_j \ , \qquad j=2,\ldots,m$$

we may vary x (which uniquely determine the $\hat{x}^j$) and solve each of the m-1 system of equations for β_j, $j = 2,\ldots,m$ provided that we select $K + 1$ linearly independent values of x not collinear with a vector of constants. Condition (iii) ensures that we can find at least $K + 1$ linearly independent vectors that are not collinear with the constant so that β_j, $\ell(j)$, $j = 2,\ldots,m$ can be uniquely identified up to scale and the $\ell(j)$, $j = 2,\ldots,m$ up to scale and location parameter. ■

Proof of Corollary to Theorem 3.

Assume $X_1 = X_2 = \ldots = X_{\bar{S}} = X$ and $\beta_1 = \beta_2 = \ldots. = \beta_{\bar{S}} = \beta$ but all other assumptions in Theorem 3 other than (v) and (vi) are valid. From (A-5-1), we may identify F_{η_1}, $\ell(1)$ and β up to normalizations using the same argument as used in the proof of Theorem 3. However, because condition (vi) no longer holds, we cannot take the second step in the proof of Theorem 3 (varying the first argument in (A-5-2) holding the second argument fixed).

Instead we know the left hand side of

$$Pr(D_1 = 1, D_2 = 1 \mid X = x) = F_{\eta_1,\eta_2}(x\beta+\ell, x\beta+\ell)$$

and cannot vary the first argument without also varying the second argument. So we can only identify F_{η_1,η_2} along the 45° diagonal in the general support $(L_1, U_1) \times (L_2, U_2)$ of (η_1, η_2).

We cannot distinguish $F(x\beta, x\beta)$ from

$$M(x\beta+\ell, x\beta+\ell) = F(x\beta+\ell, x\beta+\ell) + \chi(x\beta+\ell, x\beta+\ell)$$

where $\chi \geq 0$ and

(A-6a) $$\chi(a,a) = 0$$

(A-6b) $$\lim_{a \to \infty} \chi(a,b) = 0$$

(A-6c) $$\lim_{b \to \infty} \chi(a,b) = 0$$

(A-6d) $$\lim_{\substack{b \to -\infty \\ a \to -\infty}} \chi(a,b) = 0.$$

An example of such a χ function is

$$\chi(a,b) = K(a,b) \exp - \left[\frac{1}{(a-b)^2} + (a-b)^2 \right]$$

where

$$K(a,b) < (1 - F(a,b)) \exp \left(\frac{1}{(a-b)^2} + (a-b)^2 \right)$$

for bivariate distributions with polynomial tails.■

In this case the dependence properties of $F(\eta_1, \eta_2)$ are not identified even if we assume that η_1, η_2 are exchangeable. (So we can identify the marginals F_{η_1}, F_{η_2}). Yet it is the nonidentified dependence properties that give rise to heterogeneity bias. These results obviously extend to the general multivariate case.

Proof of Theorem 4.

First consider the probability that a person completes at least one year of schooling. To simplify notation include a constant among the regressors:

$$Pr(D_1 = 1 | X = x) = \sum_{i=1}^{I} F(x\beta_1 + \nu_i)P_i,$$

where in this proof, ν_i corresponds to the components of ν. Suppose that the model is not identified. Then there exist alternative values $\beta_1^* \neq \beta_1$, $\nu_i^* \neq \nu_i$ (for some i) and $P_i^* \neq P_i$, and $I^* \neq I$ so that so the right hand sides of the two preceding expressions are equal and

$$Pr(D_1 = 1 | X=x) = \sum_{i=1}^{I^*} F(x\beta_1^* + \nu_i^*)P_i^*$$

$$\text{(A-7)} \qquad 0 = \sum_{i=1}^{Min[I,I^*]} \left[F(x\beta_1 + \nu_i)P_i - F(x\beta_1^* + \nu_i^*)P_i^* \right] + 1[I>I^*] \sum_{i=Min[I,I^*]+1}^{I} F(x\beta_1+\nu_i)\ P_i$$
$$- 1[I^* < I] \sum_{i=Min[I,I^*]+1}^{I^*} F(x\beta_1^* + \nu_i^*)\ P_i^* .$$

If $\beta_1^* \neq \beta_1$ and they are not proportional, we can fix $x\beta_1^* = c_1$ and if $\mathcal{J}_1 = R^1$ we can set $x\beta_1 \rightarrow \pm\infty$ and hence (A-7) cannot be satisfied. Thus $\beta_1 = \beta_1^*$ (at least up to scale) if $\mathcal{J}_1 = R^1$.

Suppose $\mathcal{J}_1 \subset R_1$ is a bounded interval so that we cannot set $x\beta_1 \rightarrow \pm\infty$. Then the theorem is still true. To show this, let $X\beta_1 = \Delta$ for notational convenience. Let $x\beta_1^* = \Delta^*$. Then from Yakowitz and Spragins (1968) we know that the two sets of parameters, $\{P_i^*, \nu_i^*\}_{i=1}^{I^*}$ and $\{P_i, \nu_i\}_{i=1}^{I}$ are uniquely identified if there exist at least I values of Δ, Δ^* (Δ_i, Δ_i^* respectively) such that

$$0 \neq \begin{vmatrix} F(\Delta_1+\nu_1) & F(\Delta_1+\nu_2) & \cdots & F(\Delta_1+\nu_I) \\ \vdots & \vdots & \vdots & \vdots \\ F(\Delta_I+\nu_1) & \cdots & \cdots & F(\Delta_I+\nu_I) \end{vmatrix}$$

and the same determinantal condition holds if Δ_i^*, ν_i^* are substituted for Δ_i and ν_i. But there exist I (I^* respectively) such values by virtue of (iv).

From log-concavity of F, this determinant is positive because F is TP_2 (Karlin, 1968). Thus for β_1, β_1^* we have an associated unique $\{\nu_i, P_i\}_{i=1}^{I}$ and $\{\nu_i^*, P_i^*\}_{i=1}^{I^*}$ respectively. Now fix values of $X = x$ such that $x\beta_1^* = c_1^*$ and vary x within interval $\mathcal{J}_1$ subject to this constraint. Since $\beta_1 \neq \beta_1^*$, by hypothesis, we change the argument with β_1 on the right hand side of (A-7) but not the argument with β_1^*. Thus for (A-7) to hold $\beta_1 = \beta_1^*$. But then $\nu_i = \nu_i^*$, $P_i = P_i^*$ since given β, and the log-concavity of F we identify the mixture uniquely.

Next consider the information on the probability of completing at least two grades:

$$Pr(D_1=1, D_2=1 | X=x) = \sum_{i=1}^{I} F(x\beta_1 + \nu_i) F(x\beta_2 + \alpha_2 \nu_i) P_i.$$

Suppose that there exist alternative parameters (β_2^*, α_2^*) such that

$$Pr(D_1=1, D_2=1 | X=x) = \sum_{i=1}^{I} F(x\beta_1 + \nu_i) F(x\beta_2^* + \alpha_2^* \nu_i) P_i$$

so that

$$(A\text{-}8) \qquad 0 = \sum_{i=1}^{I} [F(x\beta_1 + \nu_i) P_i][F(x\beta_2 + \alpha_2 \nu_i) - F(x\beta_2^* + \alpha_2^* \nu_i)].$$

Then using (iv) we may set $x\beta_2^* = c_2^*$, and if we may select $x\beta_2 \to \pm\infty$, (A-8) cannot hold if $\alpha_2 = \alpha_2^* = 1$. Thus $\beta_2 = \beta_2^*$ (up to a constant of proportionality). Similarly, if we set $\beta_2 = \beta_2^*$, $\alpha_2 = \alpha_2^*$ for if $\alpha_2 > \alpha_2^*$ the expression on the right hand side of (A-8) is positive; if $\alpha_2 < \alpha_2^*$, the expression is negative.

If $\mathcal{J}_1$ is a bounded interval, we may fix x within $\mathcal{J}_1$ so that $x\beta_2^* = c_2$. Then since $\beta_2 \neq \beta_2^*$, but $\alpha_2 = 1$ we can vary the first argument $(x\beta_2)$ but not the second, and so (A-8) cannot hold over the interval. Therefore $\beta_2 = \beta_2^*$. If $\beta_2 = \beta_2^*$, by the previous argument $\alpha_2 = \alpha_2^*$.

By fixing α_s, or β_s in each transition, and continuing in this way for $s = 1,\ldots,\bar{S}$, invoking (iv), we establish Theorem 4. ■

Observe that if the x values are all discrete valued, and there are more than I distinct values, we can identify β_s only within intervals. Note further that by constraining α over some intervals we can identify different β_s. By constraining β_s we can identify different α_s.

Proof of Theorem 5.

Proceed as in the proof of Theorem 4 until expression (A-8). We get a contradiction of (A-8) since $\underset{i}{\text{Min}} |\nu_i|$, $\underset{i}{\text{Max}} |\nu_i| < \infty$ if

$$x(\beta_2 - \beta_2^*) > \underset{i=1,\ldots,I}{Max} [(\alpha_2^* - \alpha_2)\nu_i]$$

or

$$x(\beta_2 - \beta_2^*) < \underset{i=1,\ldots,I}{Min} [(\alpha_2^* - \alpha_2)\nu_i].$$

By virtue of (iv)′, we can find a value of x such that one of these inequalities is satisfied. ■

Appendix B: Data

Our analysis employs the 1973 Occupational Change in a Generation (OCG) Survey. A complete description of the data can be found in Hauser and Featherman (1975). We also use a sample of white males from the NLSY. See Cameron and Heckman (1994) for a description. The OCG data were collected as a supplement to the March 1973 Current Population Survey (CPS) and targeted the U.S. male civilian noninstitutional population aged 20 to 65. The data include information on the socioeconomic background of 37,964 men merged with the usual March CPS labor market information. Because the samples of blacks and hispanics were small, we confine our study to approximately 33,500 white males. We deleted all observations with missing values in any of the variables used in the analysis: region of birth, residence at age 16, highest grade completed of mother and father, whether the individual lived with both his parents at age 16 (broken home), parental family income at age 16, number of siblings, the individual's highest grade attended and highest grade completed, whether the individual was enrolled in school at the survey date, and occupation of the father at age 16. Occupation of the father was used to construct a Duncan's Index of the father's occupational status (Duncan, 1961). It was not used in our main results but was used to ensure comparability Mare's original analysis of schooling transitions. (See Cameron and Heckman, 1994). After deleting missing values, approximately 22,000 observations remained. We excluded individuals under age 26 because many were still in school, and the remaining 18,354 observations we divided into 4 age cohorts: ages 26-35, ages 36-45, ages 46-55, and ages 56 to 65.

The parental family income question was problematic because the survey asked individuals ranging in ages from 21 to 65 the amount of income earned by the parents in the individual's 16^{th} year. Aside from measurement error issues arising from recall uncertainty, the scale of the amounts has also been called into question. However, the original data collection group determined that a CPI deflator should be applied to this variable. Therefore, we follow Mare's 1980 analysis and deflate family income using a 3-year moving average of the CPI to construct the parental family income variable used in this paper.

The NLSY sample was constructed to mimic the OCG data as closely as possible, though some small differences in the definition of some of the family background variables were unavoidable. Highest grade completed of each parent, farm residence, and broken home were measured at age 14. Family income could not be measured at age 16 for the entire sample; the variable reported here measures family income as closely as possible to age 16. In addition family income is averaged over two years to reduce measurement error. About 12 percent of the NLSY sample was lost because the family income information was missing.

Table B.1

Sample Means and Levels of Schooling Attainment
(Standard Errors of the Mean in Parentheses)

A. Schooling Attainment

Birth Cohort (Data)	1957-1964 (NLSY)	1937-1946 (OCG)	1927-1936 (OCG)	1917-1926 (OCG)	1907-1916 (OCG)
Complete Less Than 8	0.03 (.003)	0.03 (.002)	0.07 (.004)	0.09 (.004)	0.15 (.006)
Complete 8	0.01 (.022)	0.03 (.002)	0.05 (.003)	0.09 (.004)	0.15 (.006)
Attend High School	0.11 (.007)	0.12 (.004)	0.14 (.005)	0.17 (.005)	0.19 (.007)
Graduate High School	0.30 (.010)	0.36 (.007)	0.34 (.007)	0.34 (.007)	0.27 (.008)
Attend College	0.30 (.010)	0.21 (.005)	0.16 (.005)	0.15 (.005)	0.11 (.005)
Complete College	0.13 (.007)	0.11 (.005)	0.11 (.005)	0.10 (.004)	0.07 (.004)
Attend 17 or More	0.13 (.007)	0.13 (.005)	0.12 (.005)	0.08 (.004)	0.06 (.004)
B. Means of the Independent Variables					
Number of Siblings	2.95 (.04)	3.29 (.04)	3.75 (.04)	4.10 (.04)	4.63
Family Income Age 16*	47.74 (.09)	40.68 (.08)	37.71 (.09)	36.31 (.09)	27.41 (.11)
HGC Father	11.90 (.05)	9.71 (.05)	8.31 (.06)	7.65 (.06)	7.11 (.07)
HGC Mother	12.00 (.07)	10.28 (.05)	9.08 (.05)	8.13 (.05)	7.33 (.06)
Broken Home	0.13 (.01)	0.11 (.01)	0.12 (.01)	0.13 (.01)	0.14 (.01)
Farm Residence Age 16	0.07 (.01)	0.14 (.01)	0.21 (.01)	0.27 (.01)	0.33 (.01)
Southern Birth	0.26 (.01)	0.30 (.01)	0.30 (.01)	0.28 (.01)	0.28 (.01)
Number of Observations	2098	5447	4602	4825	3480

* Family Income is denominated in 1000's of 1995 dollars.

Variable Definitions: "Family Income Age 16" is the annual income of the individual's parents. "HGC Father" and "HGC Mother" are the highest grade attained by the individual's father and mother, "Broken Home" is a binary variable indicating whether one or more of the individual's parents were absent from his household most of the time up to age 16; "Farm Residence" is an indicator recording whether the individual lived on a farm at age 16; "Southern Birth" records whether or not the individual was born in the South census region.

Table B.2
Educational Transition Probabilities for OCG (CPS 1973) and NLSY White Males
Logistic Probability Estimates (t-values)[*]

	Elementary Complete	Attend High School	Graduate High School	Attend College	Graduate College	Attend 17+
A. Born 1957-1964 (Age 26-33 in 1991-NLSY)						
Intercept	-0.228 (0.3)	2.867 (2.2)	-0.896 (2.0)	-3.577 (9.9)	-3.704 (8.6)	-1.550 (2.4)
Number of Siblings	-0.085 (1.4)	-0.192 (2.3)	-0.107 (3.0)	-0.111 (3.8)	-0.100 (2.7)	-0.207 (3.3)
Family Income	0.189 (3.4)	0.187 (2.7)	0.076 (4.2)	0.014 (1.3)	0.023 (2.0)	0.011 (0.7)
HGC Father	0.138 (1.6)	-0.053 (0.5)	0.129 (3.1)	0.129 (4.1)	0.176 (4.9)	-0.004 (0.1)
HGC Mother	0.161 (2.6)	0.181 (2.1)	0.108 (3.3)	0.230 (9.6)	0.098 (3.7)	0.145 (0.4)
Broken Home	0.481 (1.1)	-1.047 (2.3)	-0.480 (2.5)	0.189 (1.0)	-0.280 (1.0)	-0.508 (1.3)
Farm Residence	1.144 (1.5)	0.091 (0.1)	0.454 (1.5)	-0.278 (1.4)	0.005 (0.0)	-0.717 (3.3)
Southern Birth	-1.238 (4.0)	-0.956 (2.1)	-0.219 (1.3)	0.207 (1.6)	0.220 (1.5)	-0.784 (2.4)
B. Born 1927-1936 (Age 36-45 in 1973)						
Intercept	0.596(2.42)	2.152 (7.33)	0.215(1.20)	-1.212(7.85)	-0.406(1.93)	0.088(0.33)
Number of Siblings	-0.087(4.13)	-0.128 (5.89)	-0.098(6.51)	-0.082(5.75)	-0.103(4.92)	-0.013(0.44)
Family Income	0.163(9.10)	0.102(6.74)	0.072(6.91)	0.051(7.48)	0.028(3.81)	-0.004(0.51)
HGC Father	0.186(6.52)	0.068(2.32)	0.062(3.61)	0.062(4.69)	0.063(3.76)	0.003(0.13)
HGC Mother	0.117(4.54)	0.063(2.11)	0.083(4.76)	0.071(4.90)	0.012(0.60)	0.013(0.52)
Broken Home	0.226(1.09)	-0.151(0.73)	0.087(0.64)	-0.195(1.70)	-0.317(1.91)	-0.376(1.65)
Farm Residence	-0.659(4.65)	-0.960(6.49)	0.136(1.20)	-0.465(4.52)	-0.088(0.55)	0.004(0.02)
Southern Birth	-0.596(4.35)	-0.176(1.15)	-0.079(0.79)	0.004(0.05)	0.189(1.59)	-0.222(1.52)
C. Born 1917-1926 (Age 46-55 in 1973)						
Intercept	1.117(6.15)	1.453(7.33)	0.524(3.58)	-1.190(8.59)	-0.539(2.77)	-0.251(0.95)
Number of Siblings	-0.074(4.03)	-0.092(4.83)	-0.092(6.47)	-0.101(7.24)	-0.052(2.38)	-0.006(0.18)
Family Income	0.103(6.84)	0.045(4.01)	0.039(5.65)	0.044(8.41)	0.010(1.62)	-0.003(0.38)
HGC Father	0.124(5.35)	0.078(3.33)	0.076(4.64)	0.059(4.34)	0.032(1.78)	-0.018(0.79)
HGC Mother	0.166(7.47)	0.109(4.78)	0.054(3.53)	0.068(4.75)	0.052(2.65)	0.035(1.36)
Broken Home	-0.421(2.70)	-0.240(1.45)	-0.247(2.10)	-0.051(0.45)	-0.224(1.40)	-0.109(0.48)
Farm Residence	-0.527(4.29)	-1.222(10.79)	-0.108(1.10)	-0.482(4.80)	-0.303(1.95)	-0.065(0.28)
Southern Birth	-1.069(8.78)	0.411(3.31)	-0.188(1.99)	0.198(2.07)	0.022(0.17)	-0.387(2.24)
D. Born 1907-1916 (Age 56-65 in 1973)						
Intercept	1.103(6.37)	0.950(5.51)	-0.051(0.31)	-1.026(5.84)	-0.236(0.98)	-0.299 (0.90)
Number of Siblings	-0.116(6.63)	-0.115(6.77)	-0.096(5.91)	-0.075(3.94)	-0.037(1.30)	0.011(0.27)
Family Income	0.145(9.62)	0.071(5.90)	0.075(5.74)	0.055(7.15)	0.018(2.25)	-0.003(0.28)
HGC Father	0.113(4.71)	0.077(3.49)	0.054(2.77)	0.072(3.99)	0.018(0.78)	0.046(1.44)
HGC Mother	0.113(4.74)	0.088(3.85)	0.085(4.24)	0.024(1.24)	0.014(0.54)	-0.026(0.74)
Broken Home	-0.696(4.77)	-0.041(0.25)	-0.348(2.55)	0.005(0.03)	0.077(0.34)	-0.361(1.16)
Farm Residence	-0.654(5.57)	-1.184(10.90)	-0.087(0.78)	-0.364(2.88)	-0.098(0.50)	0.200(0.71)
Southern Birth	-0.769 (6.48)	0.457(3.61)	-0.075(0.65)	0.143(1.11)	0.083(0.47)	0.097(0.41)

[*]See the base of Table B.1 for variable definitions.

Table B.3

Educational Transition Probabilities for OCG and NLSY White Males
Logistic Probability Estimates with Nonparametric Heterogeneity
Correction (t-values)*

	Elementary Complete	Attend High School	Graduate High School	Attend College	Graduate College	Attend 17+
A. Born 1957-1964 (Age 26-33 in 1991-NLSY)						
Factor Leading	4.128 (8.7)	4.128 (8.7)	4.128 (8.7)	4.128 (8.7)	4.128 (8.7)	4.128 (8.7)
Intercept	-0.756 (0.9)	2.649 (2.0)	-1.363 (2.8)	-4.672 (10.1)	-6.214 (9.3)	-7.134 (5.2)
Number of Siblings	-0.084 (1.3)	-0.198 (2.3)	-0.121 (3.1)	0.142 (4.2)	-0.194 (3.7)	-0.483 (4.4)
Family Income	0.191 (3.4)	0.181 (2.6)	0.075 (4.0)	0.015 (1.3)	0.034 (2.5)	0.033 (1.5)
HGC Father	0.157 (1.8)	-0.057 (0.5)	0.144 (3.3)	0.155 (4.3)	0.229 (5.1)	0.313 (3.8)
HGC Mother	0.179 (2.8)	0.201 (2.2)	0.129 (3.7)	0.284 (10.2)	0.207 (5.4)	0.205 (2.6)
Broken Home	0.564 (1.3)	-0.999 (2.2)	-0.472 (2.4)	0.215 (1.0)	-0.438 (1.5)	-0.095 (0.2)
Farm Residence	1.175 (1.5)	0.085 (0.1)	0.394 (1.2)	-0.405 (1.8)	0.084 (0.2)	-1.149 (1.5)
Southern Birth	-1.200 (3.8)	-0.982 (2.1)	-0.244 (1.4)	0.173 (1.2)	-0.062 (0.3)	-1.115 (3.2)
B. Born 1927-1936 (Age 36-45 in 1973)						
Factor Loading	16.660 (0.0)	2.988 (9.2)	2.988 (9.2)	2.988 (9.2)	2.988(9.2)	2.988(9.2)
Intercept	0.412 (1.6)	1.916 (6.3)	-0.139 (0.7)	-1.964 (9.7)	-1.723 (5.6)	-2.054 (4.0)
Number of Siblings	-0.105 (4.6)	-0.145 (6.4)	-0.118 (7.2)	-0.127 (7.0)	-0.199 (6.6)	-0.013 (2.9)
Family Income	0.158 (8.7)	0.102 (6.7)	0.074 (7.0)	0.059 (7.9)	0.041 (4.8)	0.014 (1.2)
HGC Father	0.198 (6.6)	0.080 (2.6)	0.074 (4.1)	0.082 (5.4)	0.103 (5.1)	0.055 (1.9)
HGC Mother	0.126 (4.7)	0.074 (2.4)	0.099 (5.3)	0.104 (6.0)	0.060 (2.4)	0.083 (2.3)
Broken Home	0.180 (0.8)	-0.181 (0.9)	0.066 (0.5)	-0.235 (1.8)	-0.452 (2.1)	-0.734 (2.1)
Farm Residence	-0.698 (4.7)	-1.036 (6.8)	0.054 (0.4)	-0.668 (5.1)	-0.506 (2.3)	-0.558 (1.6)
Southern Birth	-0.542 (3.8)	-0.136 (0.9)	-0.038 (0.4)	0.035 (0.4)	0.260 (1.8)	-0.202 (1.0)
C. Born 1917-1926 (Age 46-55 in 1973)						
Factor Loading	0.782 (0.5)	2.865 (7.4)	2.865 (7.4)	2.865 (7.4)	2.865 (7.4)	2.865 (7.4)
Intercept	1.055 (5.0)	1.308 (6.3)	0.294 (1.8)	-1.855 (9.2)	-1.877 (5.4)	-2.642 (4.5)
Number of Siblings	-0.075 (4.1)	-0.100 (5.1)	-0.104 (6.8)	-0.141 (7.6)	-0.133 (4.0)	-0.102 (2.1)
Family Income	0.101 (6.6)	0.041 (3.6)	0.035 (5.0)	0.044 (7.8)	0.013 (1.8)	0.005 (0.5)
HGC Father	0.127 (5.4)	0.087 (3.6)	0.087 (5.1)	0.082 (5.1)	0.069 (3.0)	0.026 (0.8)
HGC Mother	0.166 (7.4)	0.114 (4.8)	0.063 (3.9)	0.099 (5.7)	0.117 (4.3)	0.138 (3.4)
Broken Home	-0.437 (2.8)	-0.297 (1.7)	-0.315 (2.5)	-0.183 (1.4)	-0.463 (2.3)	-0.429 (1.3)
Farm Residence	-0.531 (4.3)	-1.251(10.7)	-0.157 (1.5)	-0.669 (5.3)	-0.706 (3.1)	-0.636 (1.8)
Southern Birth	-1.066 (8.7)	0.430 (3.4)	-0.179 (1.8)	0.206 (1.8)	0.000 (0.0)	-0.511 (2.1)

Table continues on next page

D. Born 1907-1916 (Age 56-65 in 1973)						
Factor Loading	2.046 (1.0)	2.279 (5.3)	2.279 (5.3)	2.279 (5.3)	2.279 (5.3)	2.279 (5.3)
Intercept	0.975 (5.0)	0.720 (3.8)	-0.434 (2.1)	-1.759 (6.6)	-1.385 (3.3)	-2.118 (3.1)
Number of Siblings	-0.125 (6.7)	-0.127 (6.9)	-0.115 (6.3)	-0.115 (4.6)	-0.104 (2.6)	-0.055 (1.0)
Family Income	0.141 (9.1)	0.071 (5.7)	0.077 (5.5)	0.060 (7.1)	0.027 (2.9)	0.013 (1.0)
HGC Father	0.113 (4.6)	0.081 (3.5)	0.063 (3.1)	0.096 (4.5)	0.052 (1.8)	0.103 (2.4)
HGC Mother	0.126 (5.0)	0.107 (4.4)	0.111 (5.0)	0.057 (2.5)	0.055 (1.7)	0.008 (0.2)
Broken Home	-0.744 (4.8)	-0.091 (0.5)	-0.444 (3.0)	-0.115 (0.6)	-0.054 (0.2)	-0.569 (1.4)
Farm Residence	-0.681 (5.5)	-1.265(10.9)	-0.200 (1.6)	-0.575 (3.6)	-0.435 (1.6)	-0.140 (0.4)
Southern Birth	-0.774 (6.2)	0.462 (3.4)	-0.076 (0.6)	0.146 (1.0)	0.098 (0.5)	0.114 (0.4)

*See the base of Table B.1 for variable definitions.