Estimation of Dynamic Discrete Choice Models by Maximum Likelihood and the Simulated Method of Moments*

Philipp Eisenhauer
The University of Chicago

James J. Heckman
The University of Chicago

Stefano Mosso
The University of Chicago

University College Dublin
American Bar Foundation

August 14, 2013

*For further information or questions and suggestions, please contact us at info@policy-lab.org. We are indebted to George Yates for numerous valuable comments and for excellent computational assistance in developing and testing the ML estimator used in this paper and for assisting in the study of accuracy bounds for the computational algorithm. We thank Edward Sung and Jake Torcasso for their outstanding research assistance. We have benefitted greatly from comments received from Stefan Wild. This research was supported in part by the American Bar Foundation, the Pritzker Children’s Initiative, the Buffett Early Childhood Fund, NICHD 5R37HD065072, 5R01HD054702, the Human Capital and Economic Opportunity Global Working Group - an initiative of the Becker Friedman Institute for Research in Economics - funded by the Institute for New Economic Thinking (INET), and an anonymous funder. We acknowledge the support of an European Research Council grant hosted by the University College Dublin, DEVHEALTH 269874. The views expressed in this paper are those of the authors and not necessarily those of the funders or commentators mentioned here. The website for this paper is https://heckman.uchicago.edu/MLvsSMM.
Abstract

We compare the performance of maximum likelihood (ML) and simulated method of moments (SMM) for estimating dynamic discrete choice models. We construct and estimate a deliberately simplified structural model that captures some basic features of educational choices in the United States in the 1980s and early 1990s. The model is computationally tractable and amenable to exact analysis of numerical accuracy. We use estimates from our model to simulate a synthetic dataset. We assess the ability of ML and SMM to recover the model parameters on this dataset. We conduct an analysis of the numerical precision of the ML estimator and present precise error bounds for our computational algorithm. ML estimates applied to the synthetic data are close to the true parameter values, while conventional formulations of SMM have trouble recovering some key model parameters. We investigate the performance of alternative tuning parameters for SMM. We show how the choice of moments, number of replications, weighting matrix, and optimization algorithm affects the quality of estimates obtained from SMM.

**JEL Codes:** C53, C52, C32, C38

**Key Words:** Dynamic Discrete Choice, Maximum Likelihood, Simulated Method of Moments
1 Introduction

Economic science uses economic theory to guide the interpretation of economic data and to shape policy. Kenneth Wolpin is a model economic scientist who integrates theory and data in a rigorous fashion. He is a major contributor to structural econometrics with particular emphasis on the study of dynamic discrete choice models. His contributions are both methodological and empirical. His methodological research focuses on promoting methods to increase the reliability of structural estimation algorithms [Eckstein and Wolpin 1989; Keane et al. 2010] and to develop techniques to simplify their empirical implementation. His research on interpolation methods to solve dynamic discrete choice models with a large state space [Keane and Wolpin 1994] is one prominent example. In his empirical contributions, he extensively applies these methods to investigate many important issues such as educational attainment [Eckstein and Wolpin 1999; Keane and Wolpin 1997], the role of credit constraints in educational attainment [Keane and Wolpin 2001], and labor market dynamics [Lee and Wolpin 2006 2010].

This paper contributes to the literature on estimating dynamic discrete choice models. It investigates the empirical performance of widely used versions of simulated method of moments (SMM), a tractable method for estimating complex structural models. SMM estimates parameters by fitting a vector of empirical moments to their theoretical counterparts simulated from a structural model [McFadden 1989].

We estimate a dynamic discrete choice model of schooling based on a sample of white males from the National Longitudinal Survey of Youth (1979) with some key features in common with the model of [Keane and Wolpin 1997]. We greatly simplify the model in order to characterize the likelihood function analytically. Its evaluation does not require simulation. This provides a clean comparison of ML against simulation-based estimation methods such as SMM. It allows us to analytically characterize the errors of approximation in constructing the likelihood.

In a series of Monte Carlo simulation studies we compare estimates based on our precisely calculated ML with those from widely used, computationally tractable versions of SMM. Using the estimates of model parameters, we simulate a dataset on which we compare ML and SMM estimation.
Because our synthetic sample is derived from real data, our analysis provides useful lessons on the performance of SMM for the estimation of structural models.

SMM has been used to estimate models of job search (Flinn and Mabli 2008), educational and occupational choices (Adda et al. 2013, 2011), household choices (Flinn and Del Boca 2012), stochastic volatility models (Andersen et al. 2002; Raknerud and Skare 2012), and dynamic stochastic general equilibrium models (Ruge-Murcia 2012). SMM can be used for any model, however complex or difficult to compute its likelihood function, as long as it is possible to simulate it. Under conditions presented in the literature, the SMM estimator is consistent and asymptotically normal. If the score vector for SMM spans the space of the score vectors of the likelihood, SMM is asymptotically efficient (Gach 2010; Gach and Pötscher 2011; Nickl and Pötscher 2009).

The performance of SMM has rarely been evaluated. Most papers that use SMM for estimation do not offer any study of its reliability. There are many choices available to the user of SMM and the consequences of these choices are rarely studied. The criteria often employed to evaluate model fit in SMM studies are quite weak in assessing its performance. Implementing the SMM estimator requires numerous, necessarily arbitrary choices. Users have discretion in selecting: (1) the moments used in estimation, (2) the number of replications used to compute the simulated moments, (3) the moment weighting matrix, and (4) the algorithm used for optimization. It is unclear how such choices affect the performance of the SMM estimator and how they depend on the structure of the model estimated. We propose diagnostic tools to test their validity. We show how application of more exacting goodness of fit criteria can point to potential model weaknesses.

Section 2 presents our schooling model. Section 3 presents baseline results on the returns to school-

---

1 Alternative estimation methods have been proposed to overcome the rigidities and complexities of ML estimation. Most require the analyst to characterize the likelihood function, but simplify its computation. One of the most popular methods, simulated ML (SML), substitutes the exact likelihood function with a simulated one. An example is the Hajivassiliou-Geweke-Keane (HGK-SML) estimator (Geweke 1989; Hajivassiliou and McFadden 1998; Keane 1994) used for multinomial probit estimation. Approximations of the dynamic programming problem have often been combined with SML in models with a large state space (Keane and Wolpin 1994). Another popular method is the conditional choice probabilities (CCP) algorithm first proposed by Hotz and Miller (1995) and recently extended to allow for unobserved heterogeneity (Arcidiacono and Miller 2011). Using CCP, a consistent estimator of the model parameters can be derived without the need of the full solution of the dynamic programming problem. The CCP method, however, restricts the flexibility of the estimable models by imposing assumptions which limit the dependence among successive choices. SMM is a more general alternative to ML estimation.
ing. Section 4 outlines our Monte Carlo study and compares the performance of ML and SMM estimation. Section 5 concludes.

2 Dynamic Model of Educational Choices

We first present a dynamic discrete choice model of education and establish its identification. We specify a model with a simple state space by assuming that agents move from one schooling state to the next. We restrict agents to binomial choices at each decision node. The agents optimally choose between two alternatives at each node. The value of each state is determined by its immediate earnings and costs, and by the expected future value of all feasible states. Agents have private information on their own type and form expectations about future states with respect to their current information set.

Our simplifications come at the expense of a less realistic empirical analysis of the dynamics of schooling choices compared to those of Keane and Wolpin (1997, 2001) and Johnson (2012). We make these assumptions because they allow us to exactly characterize the error in computing the likelihood and evaluating it does not require any simulation. Our likelihood function can be computed at a very high level of numerical precision. We are able to calculate the terms in the likelihood with accuracy up to 15 decimal places. This guarantees at least three digits of accuracy for all estimated model parameters. We discuss the numerical properties of the likelihood and bounds on approximation error in Web Appendix A.

2.1 Setup

Given the current state \( s \in S = \{s_1, \ldots, s_N\} \), let \( S^v(s) \subseteq S \) denote the set of visited states and \( S^f(s) \subseteq S \) the set of feasible states that can be reached from \( s \). We collect the choice set of the agent in state \( s \) in \( \Omega(s) = \{s' | s' \in S^f(s)\} \). We consider binary choices only, so \( \Omega(s) \) has at most two elements. Ex post, the agent receives per period rewards \( R(s') = Y(s') - C(s', s) \) defined as the difference between per period earnings, \( Y(s') \) and the costs \( C(s', s) \) associated with moving from state \( s \) to state \( s' \). The cost of one of the exits is normalized to zero. We collect the subset of states with a costly exit in \( S^c \). In the subsequent analysis it is useful to explicitly distinguish between the nonzero (\( \hat{s}' \)) and zero
cost \((s')\) exits from \(s\). Earnings and costs have deterministic and stochastic components. We assume that they are separable functions of observed covariates \(X(s) \in \mathcal{X}\) for outcomes and \(Q(s', s) \in \mathcal{Q}\) for costs. In addition, earnings and costs depend on individual specific factors \(\theta\) that are known to agents but unknown to the econometrician and idiosyncratic shocks \(\epsilon(s)\) and \(\eta(s', s)\) that are unknown to the econometrician and only partly known by the agent. We allow idiosyncratic shocks to be independent, but not identically distributed. We thus generalize the i.i.d. innovation assumption in Keane and Wolpin (1997). Earnings are expressed as:

\[
Y(s) = \mu_s(X(s)) + \theta' \alpha_s + \epsilon(s) \tag{1}
\]

The costs of going from state \(s\) to feasible state \(s'\) are defined by:

\[
C(s', s) = \begin{cases} 
K_{s', s}(Q(s', s)) + \theta' \phi_{s', s} + \eta(s', s) & \text{if } s' = \hat{s}' \\
0 & \text{if } s' = \tilde{s}'. 
\end{cases} \tag{2}
\]

The vectors of covariates \(Q(s', s)\) and \(X(s')\) might have elements in common. Their distinct elements constitute the exclusion restrictions.

Following Carneiro et al. (2003), Cunha et al. (2010), and Heckman et al. (2013b), we assume access to a \(J\) dimensional vector of individual measures \(M\) (such as test scores or behavioral indicators) proxying individual factors \(\theta\). We use the measures as noisy signals of the factors \(\theta\):

\[
M(j) = \mu_j(X(j)) + \theta' \gamma_j + \nu(j) \quad \text{for } j = 1, \ldots, J. \tag{3}
\]

We assume that \(\epsilon(s), \eta(s', s),\) and \(\nu(j)\) are mutually independent for all \(j, s, s'\). In measurement system (3), we interpret the unobserved factors as individual specific traits. The impact of these traits on earnings, measurements, and costs is given by the factor loadings \((\alpha_s, \gamma_j, \phi_{s', s})\). We allow for unobservable correlations in outcomes and choices across states through \(\theta\) and allow the loadings to vary across states.

Following Keane and Wolpin (1997), we assume that agents are risk neutral and maximize discounted
lifetime net earnings when making their educational choices. When an agent makes his educational choice to proceed from state $s$ to $s'$, he knows the stochastic component of the transition $\eta(s', s)$ but not of future earnings $\epsilon(s')$. We can represent the timing of the arrival of information in the problem as follows:

\[
\begin{align*}
&s \quad \text{realized} \quad \epsilon(s') \\
&s' \quad \text{received} \quad Y(s') \quad \text{paid} \quad C(s', s)
\end{align*}
\]

\[
\begin{align*}
\eta(s', s) \quad \text{realized} \quad s'' \in \Omega(s') \quad \text{picked} \quad \text{optimally}
\end{align*}
\]

The agents know $X(s), Q(s', s)$, and $\theta$ for all $s$. Under this timeline, we define $I(s)$ as the information set of the agent in state $s$ by specifying all components known in the $s$ state:

\[
\begin{align*}
\text{for all } s \in S^s(s) & \quad \eta(s', s); \epsilon(s) \\
\text{for } s' \in S^f(s) & \quad \eta(s', s) \\
\text{and for all } s & \quad X(s); Q(s', s); \theta
\end{align*}
\]

$\in I(s)$.  

The agents in state $s$ know the costs associated with a transition to any feasible state $s'$. We assume that the agent has rational expectations and uses the actual distributions for the earnings shocks $\epsilon(s)$, denoted by $F_{E,s}(\epsilon(s))$, and for the transition costs shock $\eta(s', s)$, denoted by $F_{H,s}(\eta(s', s))$, to form expectations of future states. We allow the distributions of the shocks to vary across states.\(^2\)

\(^2\)We differ from Keane and Wolpin (1997) in our specification of the distribution of the unobserved components. In their specification, agents have different initial conditions for each state variable. The distribution of initial conditions is multinomial with five components. They assume that there are only four types (values) of initial conditions in the population. Serial dependence is induced through the persistence of the initial conditions as determinants of current state variables. In addition, at each age the agent receives five shocks associated with the rewards of each choice. The shocks are joint normally distributed, serially uncorrelated, and they are assumed to be i.i.d. over time.

In our model, we allow for state dependence in the distribution of the unobservables by letting earnings and cost shocks be drawn from normal distributions with different variances at each state and at each transition. Moreover, we allow unobserved portions of cost and return functions to be contemporaneously and serially correlated through their common dependence on the factors $\theta$. Our $\theta$ are normally distributed so we have a continuum of types. The Keane and Wolpin (1997) specification of persistent heterogeneity is a version of a factor model in which all factor loadings are implicitly determined (through Bellman iterations) by the parameters of the deterministic portions of cost and return functions and the distribution functions of unobserved variables and the sample distribution of observables. In our approach, the factor loadings are specified independently of the parameters of the deterministic portions of the cost and return functions and the sample distribution of observed variables.
We define the agent’s value function at state $s$ recursively as:

$$V(s \mid I(s)) = Y(s) + \max_{s' \in \Omega(s)} \left\{ \frac{1}{1 + r} \left( -C(s', s) + \mathbb{E}[V(s' \mid I(s')) \mid I(s)] \right) \right\}. \quad (4)$$

For future reference we define the continuation value of state $s$ as the second term on the right hand side of (4):

$$CV(s \mid I(s)) = \max_{s' \in \Omega(s)} \left\{ \frac{1}{1 + r} \left( -C(s', s) + \mathbb{E}[V(s' \mid I(s')) \mid I(s)] \right) \right\}. \quad (5)$$

The agent’s policy function determines the optimal transitions. An agent in $s$ chooses his next feasible state $s'$ according to the following rule:

$$s' = \begin{cases} 
\hat{s}' & \text{if } \mathbb{E} \left[ V(\hat{s}') \mid I(s) \right] - C(\hat{s}', s) > \mathbb{E} \left[ V(\tilde{s}') \mid I(s) \right] \\
\tilde{s}' & \text{otherwise.} \end{cases}$$

$$\quad (6)$$

We now define the returns to schooling and the concept of the option value for attending school.

### 2.2 Returns to Education

We define the ex ante and ex post net returns to schooling. The net return ($NR$) to schooling includes per period earnings and costs associated with each educational choice and the option value of future opportunities (discussed in the next subsection). The ex ante net returns are defined before the unobservable components of future earnings are realized. They depend on agents’ expectations and determine their choices. Standard return concepts such as Mincer coefficients or internal rates of returns ignore costs and option values of future opportunities. They are only interpretable for terminal choices and ex post realized earnings streams. We define the ex ante net return of $\hat{s}'$ over $\tilde{s}'$ for an agent in state $s$ as:

$$\frac{\mathbb{E} \left[ V(\hat{s}') - V(\tilde{s}') \mid I(s) \right] - C(\hat{s}', s)}{\mathbb{E} \left[ V(\hat{s}') \mid I(s) \right]} = NR^a(\hat{s}', \tilde{s}', s).$$

$^3$See Heckman et al. (2006a) for a discussion of conventional methods for estimating rates of return and their economic interpretation.
We also define the *ex ante* gross return (GR) which includes all future earnings, but omits all psychic costs related to educational choices. Define the gross value of experiencing feasible state $s'$ after state $s$ as $\bar{V}(s' | \mathcal{I}(s')) = V(s' | \mathcal{I}(s')) + C(s', s)$ and with the following recursive structure:

\[
\bar{V}(s' | \mathcal{I}(s')) = Y(s') + \left\{ \frac{1}{1 + r} \mathbb{E}_{s'' \in \Omega(s')} [\bar{V}(s'' | \mathcal{I}(s'')) | \mathcal{I}(s')] \right\},
\]

where state $s'' \in \Omega(s')$ maximizes the discounted future earnings according to the policy function defined in equation (6). Although agents do not base their educational choices upon the gross returns, they are important as they are defined in terms of earnings streams only and are the focus of much applied work. We define the *ex ante* gross return of $\hat{s}'$ over $\tilde{s}'$ for an agent in $s$ as:

\[
\frac{\mathbb{E} [\bar{V}(\hat{s}') - \bar{V}(\tilde{s}') | \mathcal{I}(s)]}{\mathbb{E} [\bar{V}(\tilde{s}' | \mathcal{I}(s)]} = GR^a(\hat{s}', \tilde{s}', s). \tag{8}
\]

We formulate the net and gross *ex post* returns in the same way, but use the value functions which include the realizations of the imminent earnings shock. The *ex post* returns can be used to evaluate an agent’s regret of his educational choice.

### 2.3 Option Values of Schooling

Weisbrod (1962) was the first to analyze option values in the context of schooling and human capital accumulation. Consider a high school enrollee, who is contemplating to either graduate or drop out. Part of his evaluation of high school graduation is the option to start a college education in the future. From the perspective of state $s$ the option value of $s'$ is defined as the difference between the value of the possibility of taking the optimal choice when moving from $s'$ and the fallback value the agent obtains by choosing the zero cost exit $\tilde{s}''$ from $\Omega(s')$. The zero cost exit is usually associated with maintaining the current education level, e.g., remaining a high school graduate and not enrolling in college. Then the option value of feasible state $s'$ from the perspective of $s$ is:

\[
OV(s', s) = \]
\[
\frac{1}{1+r} \mathbb{E} \left[ \max_{s'' \in \Omega(s')} \left\{ -C(s'',s') + \mathbb{E} \left( V(s'') \right) \right\} - \mathbb{E} \left( V(s'') \right) \right] I(s).
\]

We define the option value contribution \( OVC(s',s) = \frac{OVC(s',s \mid I(s))}{\mathbb{E}[V(s') \mid I(s)]} \) as the relative share of the option value in the overall value of a state. This is a measure of misspecification in models where returns are computed excluding option values such as Mincer coefficients or internal rates of returns.

### 2.4 Identification

Our model is semi-parametrically identified by an extension of the arguments in [Heckman and Navarro (2007)](#). The main arguments of the proof, presented in Web Appendix B, consist of using: (1) a limit set argument to identify the joint distribution of earnings and measurements free of selection (“identification at infinity”), (2) the measurement system on the factor structure that facilitates identification of the joint distribution of factors, (3) the choice structure and exclusion restrictions to identify the distribution of costs in the last choice equation, and (4) backward induction to identify relevant distributions in all states showing that the future value function acts as an exclusion restriction in current choices. We can identify all of the parameters of the model including the discount rate.

### 3 Baseline Estimates

We fit the model on a sample of 1,418 white males from the National Longitudinal Survey of Youth of 1979 (NLSY79) using ML estimation. Figure 1 shows the decision tree for our model.

All agents start in high school and decide to either drop out or finish. If they finish high school, they can enroll in college immediately or remain high school graduates with the option to enroll in college later or not at all. Conditional on early or late college enrollment, agents can either graduate or drop out. At each decision node, we designate the lower transition to be the zero cost exit.

---

In all states $s$, agents work in the labor market and receive earnings $Y(s)$. When agents pursue higher education by transitioning to the costly state $s'$ they incur cost $C(s', s)$. Agents face uncertainty about components of future benefits and costs when determining the ex ante value of each state $V(s | \mathcal{I}(s))$ given the information $\mathcal{I}(s)$ available to them. As noted in Section 2, we assume that the agent knows his type and all past, present, and future covariates. His expectations about the distributions of all future shocks are assumed to be consistent with their actual realizations, as in standard rational expectations models. In our empirical specification, all shocks follow a normal distribution.

Following Carneiro et al. (2003) and Heckman et al. (2006b) we assume that the agent’s type $\theta$ is summarized by cognitive and non-cognitive abilities. We use the scores on the Armed Services Vocational Aptitude Battery (ASVAB) as noisy measures on cognitive abilities. For non-cognitive skills, we rely on Rotter and Rosenberg 1980 scores and indicators of risky behaviors such as drug and alcohol use.

We collapse the agent’s state experience into one event. In a state $s$, we assign each agent a duration $D(s)$ based on the number of periods spent in that state. For an agent who spent four years in college, the duration of the college enrollment state will be four. We set the duration for an agent’s counterfac-
tual state to the median duration among the agents who actually visit that state. Let $Y(t,s)$ denote the observed earnings in the NLSY79 at time $t$ for an agent in state $s$. We collapse all $Y(t,s)$ within state $s$ into one discounted average:

$$Y(s) = \frac{\sum_{t=1}^{D(s)} \left( \frac{1}{1+r} \right)^{t-1} Y(t,s)}{\sum_{t=1}^{D(s)} \left( \frac{1}{1+r} \right)^{t-1}}.$$ 

We do the same for time varying covariates in $X(s)$ and $Q(s',s)$. We do not estimate the discount factor $r$ and instead set $r = 0.04$.

We discuss the construction of our sample in Web Appendix C. The NLSY79 only has data up to approximately age 45. We extend the duration of the terminal states up to age 65 using parameters estimated on the available sample to project earnings in unobserved years. The high school enrollment state characterizes initial conditions in our model. We assume earnings and costs are functions of standard individual characteristics and local economic conditions.

Figure 1 presents the average annual earnings and the number of observations by state. Earnings are low during high school enrollment ($2,474) and during the year of graduation ($7,747). High school graduates earn ($42,919) almost twice as much as high school dropouts ($22,878). Our distinction between early and late college enrollment is important. Early enrollees earn much less while in college ($11,781) compared to late enrollees ($27,192). Also, early college graduation boosts average annual earnings to $74,646 compared to only $48,408 for late graduation. In the case of late college enrollment, the difference in earnings among graduates and dropouts is minor: $48,408 compared to $48,866. This


[6] In each state, earnings depend on the number of children in the household, parental education (as the maximum between the mother’s and father’s education), indicators for the presence of a baby (child less than 3 years old) in the household, marriage status, urban residence at age 14, the region of residence (North East, North Central, South and West), hourly wage and unemployment levels in the state of residence for the relevant age group (we use two age groups, younger than 30 years old or older). For the cost equations we exclude the indicator for marriage and the regional dummies, adding instead an indicator for whether the family is intact or not, the number of siblings, and state level tuitions for public two- and four-year colleges for the transitions to college enrollment states. The state representing the conclusion of high school is estimated using only an intercept, the two factors and an unobservable component. All transition and outcome equations also include the cognitive and non-cognitive factor and an idiosyncratic unobserved component. All parameters in the states referring to college enrollment, college graduation, and college dropouts and in the relative transition in the lower branch of the tree in Figure 1 are constrained to equal the parameters of the same states in the upper branch. We impose this constraint because there are too few agents in the lower branch (late enrollment) to estimate a separate set of parameters for these states. Hence we assume that the hedonic price of observable and unobservable characteristics is the same conditional on college enrollment or graduation, independently of the timing of such decisions.
explains why in our sample the number of late college dropouts (95) is actually larger than of late college graduates (77). For the case of early enrollment (589), the vast majority graduates (471). The Mincer coefficient is 0.116\[^7\].

### 3.1 Model Fit

Table 1 shows the fit of the model estimated by ML for model fit statistics that are typically used in the literature. Average earnings and state frequencies are well fit by our model. Small discrepancies show up for terminal states. Terminal states are populated by very few agents, which requires us to constrain the outcome and cost parameters of terminal college states to be equal for early and late college graduates.

**Table 1: Cross Section Model Fit**

<table>
<thead>
<tr>
<th>State</th>
<th>Average Earnings</th>
<th>State Frequencies</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Observed</td>
<td>ML</td>
</tr>
<tr>
<td>High School Dropout</td>
<td>2.29</td>
<td>2.59</td>
</tr>
<tr>
<td>High School Finishing</td>
<td>0.77</td>
<td>0.78</td>
</tr>
<tr>
<td>High School Graduation</td>
<td>2.72</td>
<td>2.52</td>
</tr>
<tr>
<td>Early College Enrollment</td>
<td>1.18</td>
<td>1.14</td>
</tr>
<tr>
<td>High School Graduation (cont’d)</td>
<td>4.29</td>
<td>3.84</td>
</tr>
<tr>
<td>Late College Enrollment</td>
<td>2.72</td>
<td>2.52</td>
</tr>
<tr>
<td>Early College Dropout</td>
<td>4.55</td>
<td>3.87</td>
</tr>
<tr>
<td>Early College Graduation</td>
<td>6.73</td>
<td>7.46</td>
</tr>
<tr>
<td>Late College Dropout</td>
<td>4.89</td>
<td>4.88</td>
</tr>
<tr>
<td>Late College Graduation</td>
<td>4.84</td>
<td>6.22</td>
</tr>
</tbody>
</table>

**Notes:** Earnings are discounted using the within state duration and measured in units of $10,000. Statistics are calculated on the NLSY79 sample and for ML based on 50,000 simulated agents using the parameter estimates. State frequencies are conditional on factually visiting the previous state.

\[^7\]Web Appendix D presents additional descriptive statistics and estimates of conventional internal rates of return.
Comparing the fit of the model to cross section moments is a weak criterion for a dynamic model. A more exacting criterion is to predict sequences of educational choices (Heckman, 1981). We follow Heckman and Walker (1990) and Heckman (1984) and use $\chi^2$ goodness of fit tests to examine our model’s performance. In Table 2 we report the $p$-value of a joint test of the relative share of agents for each state for all realizations of selected covariates. For most cells the fit is good.

Table 2: Conditional Model Fit

<table>
<thead>
<tr>
<th>State</th>
<th>Number of Children</th>
<th>Baby in Household</th>
<th>Parental Education</th>
<th>Broken Home</th>
</tr>
</thead>
<tbody>
<tr>
<td>High School Dropout</td>
<td>0.77</td>
<td>0.26</td>
<td>0.37</td>
<td>0.03</td>
</tr>
<tr>
<td>High School Finishing</td>
<td>0.88</td>
<td>0.73</td>
<td>0.55</td>
<td>0.35</td>
</tr>
<tr>
<td>High School Graduation</td>
<td>0.91</td>
<td>0.94</td>
<td>0.65</td>
<td>0.91</td>
</tr>
<tr>
<td>High School Graduation (cont’d)</td>
<td>0.95</td>
<td>0.33</td>
<td>0.40</td>
<td>0.85</td>
</tr>
<tr>
<td>Early College Enrollment</td>
<td>0.46</td>
<td>0.54</td>
<td>0.01</td>
<td>0.15</td>
</tr>
<tr>
<td>Early College Graduation</td>
<td>0.06</td>
<td>0.86</td>
<td>0.00</td>
<td>0.14</td>
</tr>
<tr>
<td>Early College Dropout</td>
<td>0.33</td>
<td>0.27</td>
<td>0.54</td>
<td>0.75</td>
</tr>
<tr>
<td>Late College Enrollment</td>
<td>0.80</td>
<td>0.23</td>
<td>0.90</td>
<td>0.60</td>
</tr>
<tr>
<td>Late College Graduation</td>
<td>0.90</td>
<td>0.39</td>
<td>0.90</td>
<td>0.60</td>
</tr>
<tr>
<td>Late College Dropout</td>
<td>0.89</td>
<td>0.42</td>
<td>0.91</td>
<td>0.76</td>
</tr>
</tbody>
</table>

One exception (at a 5% significance level) is Parental Education, where we fail to fit the observed patterns for early college enrollment and early college graduation. For Broken Home, we overpredict the relative share of individuals from a broken home among high school dropouts. For all other variables and states, the $p$-values indicate that the model is consistent with the data. Because tests within covariates across all states are not independent, we use a Bonferroni test to evaluate the joint hypothesis that the predicted covariate distributions fit at each state. The test is based on the maximum $\chi^2$ statistic over all states for each covariate. A 5% Bonferroni test is passed by all covariates besides Parental Education. Here, the poor prediction for early college graduates leads to an overall rejection.

---

8In the $\chi^2$ test, the predicted covariate distributions depend on estimated parameters. We do not adjust the test statistic to account for parameter estimation error as suggested by Heckman (1984) because the adjustments are usually slight (Heckman and Walker, 1990).
3.2 Economic Implications

We now present the economic implications of our baseline results. We first discuss the impact of unobserved abilities on educational choices and outcomes and then turn to the role of psychic costs and option values for the net returns to schooling.

3.2.1 Impact of Abilities

Figure 2 plots the distribution of cognitive and non-cognitive abilities for agents in each of the terminal states. Figure 3 shows the share of agents in each of the final states by deciles of the overall factor distribution.

![Figure 2: Ability Distributions by Terminal States](image)

Notes: We simulate a sample of 50,000 agents based on the estimates of the model.

The distributions of abilities differ substantially across schooling states. Early college graduates are strong in cognitive and non-cognitive abilities. High school dropouts are weak in both. High school graduates, who never enroll in college, are weak in cognitive abilities, but quite strong in non-cognitive abilities.

Figure 4 shows the transition probabilities to each state by factor deciles. Higher cognitive skills increase the likelihood of continued educational achievement for all choices. The effect of non-cognitive
Figure 3: Ability Distributions by Final Education

(a) Cognitive

(b) Non-Cognitive

Notes: We simulate a sample of 50,000 agents based on the estimates of the model.

abilities is mixed. While they clearly increase the likelihood of finishing high school, higher non-cognitive skills decrease the probability of late college enrollment (conditional on working after high school graduation). Delay of college enrollment is associated with lower levels of non-cognitive skills.

3.2.2 Psychic Costs of Educational Choices

We estimate the overall costs associated with each educational choice. Our estimated costs combine monetary expenses such as tuition and psychic costs (Cunha et al., 2005). Table 3 reports the average costs associated with each transition. It reports the second, fifth, and eighth decile of their distribution to document their substantial heterogeneity. Costs are key components of the net returns, ignoring them results in biased estimates. The largest costs are associated with early and late college enrollment. These are the only states with psychic as well as monetary costs from tuition. Enrolling early costs the
equivalent of $274,000 compared to $553,000 for late enrollment. At least 20% of agents have negative schooling costs in most states. They experience psychic benefits. For high school graduation, the average cost is negative.

Table 3: Psychic Costs

<table>
<thead>
<tr>
<th>State</th>
<th>Mean</th>
<th>2\textsuperscript{nd} Decile</th>
<th>5\textsuperscript{th} Decile</th>
<th>8\textsuperscript{th} Decile</th>
</tr>
</thead>
<tbody>
<tr>
<td>High School Finishing</td>
<td>-2.39</td>
<td>-5.55</td>
<td>-2.40</td>
<td>0.79</td>
</tr>
<tr>
<td>Early College Enrollment</td>
<td>2.74</td>
<td>-0.64</td>
<td>2.70</td>
<td>6.09</td>
</tr>
<tr>
<td>Early College Graduation</td>
<td>1.78</td>
<td>-3.98</td>
<td>1.86</td>
<td>7.63</td>
</tr>
<tr>
<td>Late College Enrollment</td>
<td>5.53</td>
<td>1.75</td>
<td>5.48</td>
<td>9.33</td>
</tr>
<tr>
<td>Late College Graduation</td>
<td>1.29</td>
<td>-4.79</td>
<td>1.45</td>
<td>7.40</td>
</tr>
</tbody>
</table>

Notes: We simulate a sample of 50,000 agents based on the estimates of the model. We condition on the agents that actually visit the relevant decision state. Costs are in units of $100,000.

3.2.3 Returns to Education

Figure 5 presents the \textit{ex ante} net return to schooling by factor deciles. The effect of latent skills on returns differs by state. The return of finishing high school is strongly affected by the non-cognitive factor. Cognitive and non-cognitive abilities are equally important for increasing the return of early college enrollment, while early college graduation is influenced by cognitive skills alone. The returns to late college enrollment are low regardless of skill level. Our estimates show evidence of strong complementarity between abilities and schooling for most states. Figure 5 also presents median returns. The median net return for early college enrollment is around zero and the return of delayed enrollment even negative (-23%). College dropouts pay the cost of college without benefiting from the much larger returns of graduating. Among those who enroll late, the returns from graduating (15%) are smaller than for those enrolling early (57%). We report the difference between net and gross returns as part of Figure 5. Psychic costs are crucial determinants of net returns. While the median gross return for early and late college enrollment is positive, the net return is negative in both cases.
Ex ante and ex post returns do not necessarily agree because agents cannot predict their future earnings. Decisions that are optimal for an agent ex ante might be suboptimal ex post. For this reason, we calculate the percentage of agents experiencing regret, i.e., those agents for whom the ex post and ex ante returns do not agree in sign. A substantial share of late college enrollees (34%) regret the decision to graduate. For finishing high school, the share is much smaller (4%). However, 24% of high school dropouts regret their decision.

3.2.4 Option Values of Schooling

Our structural model allows us to calculate the option values of educational choices. The option value is the difference in the value associated with the optimal continuation of choices versus the fallback value. Figure 6 shows the option values conditional on the deciles of the factor distributions, their median (OV), and their contribution to the total value of each state (OVC). The option values make a sizable contribution to the overall value of the states and vary by abilities.

Early college enrollment has the highest option value as graduation yields a large gain in earnings compared to dropping out. As the net returns to college graduation increase in cognitive and non-cognitive abilities, so does the option value of college enrollment.

See Web Appendix D for additional results on ex post returns and regret.

Other models taking into account option values have been proposed by Comay et al. (1973), Cunha et al. (2007), and Heckman et al. (2013a). See also Cameron and Heckman (1993).
Figure 4: Transition Probabilities by Abilities

(a) High School Finishing  (b) Early College Enrollment  (c) Early College Graduation

(d) Late College Enrollment  (e) Late College Graduation

Notes: We simulate a sample of 50,000 agents based on the estimates of the model. In each subfigure, we condition on the agents that actually visit the relevant decision state.
Figure 5: Ex Ante Net Returns by Abilities

(a) High School Finishing
NR\(^a\) = 0.64
GR\(^a\) = 0.30

(b) Early College Enrl.
NR\(^a\) = -0.06
GR\(^a\) = 0.17

(c) Early College Grad.
NR\(^a\) = 0.57
GR\(^a\) = 0.89

(d) Late College Enrl.
NR\(^a\) = -0.23
GR\(^a\) = 0.34

(e) Late College Grad.
NR\(^a\) = 0.15
GR\(^a\) = 0.33

Notes: We simulate a sample of 50,000 agents based on the estimates of the model. In each subfigure, we condition on the agents that actually visit the relevant decision state. Enrl. = Enrollment, Grad. = Graduation.
Figure 6: Option Values by Abilities

(a) High School Finishing
   \( OV = 0.99 \)
   \( OVC = 0.10 \)

(b) Early College Enrollment
   \( OV = 3.33 \)
   \( OVC = 0.30 \)

(c) Late College Enrollment
   \( OV = 2.19 \)
   \( OVC = 0.19 \)

Notes: We simulate a sample of 50,000 agents based on the estimates of the model. In each subfigure, we condition on the agents that actually visit the relevant decision state. In units of $100,000.
4 Comparison of ML and SMM

We use the baseline estimates of our structural parameters to simulate a synthetic sample of 5,000 agents. Thus our sample captures important aspects (summarized in Section 3) of our original data such as model complexity and sizable unobserved variation in agent behaviors. Then we disregard our knowledge about the true structural parameters. We estimate the model on the synthetic sample by ML and SMM to compare their performance in recovering the true structural objects. We first describe the implementation of both estimation procedures. Then we compare their within-sample model fit and assess the accuracy of the estimated returns to education. Finally, we explore the sensitivity of our SMM results to alternative tuning parameters such as choice of the moments, number of replications, weighting matrix, and optimization algorithm.

We assume the same functional forms and distributions of unobservables for ML and SMM. Measurement, outcome, and cost equations (1) - (3) are linear-in-parameters. Recall that $S^c$ denotes the subset of states with a costly exit.

$M(j) = X(j)\beta_j + \theta' \gamma_j + \nu(j)$ \hspace{1cm} $\forall$ \hspace{0.2cm} $j \in M$

$Y(s) = X(s)\beta_s + \theta' \alpha_s + \epsilon(s)$ \hspace{1cm} $\forall$ \hspace{0.2cm} $s \in S$

$C(s',s) = Q(s',s)\delta_{s',s} + \theta' \phi_{s',s} + \eta(s',s)$ \hspace{1cm} $\forall$ \hspace{0.2cm} $s \in S^c$

All unobservables of the model are normally distributed,

$\eta(s',s) \sim N(0,\sigma_{\eta(s',s)})$ \hspace{1cm} $\forall$ \hspace{0.2cm} $s \in S^c$

$\epsilon(s) \sim N(0,\sigma_{\epsilon(s)})$ \hspace{1cm} $\forall$ \hspace{0.2cm} $s \in S$

$\theta \sim N(0,\sigma_{\theta})$ \hspace{1cm} $\forall$ \hspace{0.2cm} $\theta \in \Theta$

$\nu(j) \sim N(0,\sigma_{\nu(j)})$ \hspace{1cm} $\forall$ \hspace{0.2cm} $j \in M$.

The unobservables $(\epsilon(s), \eta(s',s), \nu(j))$ are independent across states and measures. The two factors $\theta$ are independently distributed. We allow for unobservable correlations in outcomes and choices through the factor components $\theta$ (Cunha et al., 2005).

4.1 The ML Approach

We now describe the likelihood function, its implementation, and the optimization procedure.
For each agent we define an indicator function $G(s)$ that takes value one if the agent visits state $s$. Let $\psi \in \Psi$ denote a vector of structural parameters and $\Gamma$ the subset of states visited by agent $i$. We collect in $D = \{ \{ X(j) \}_{j \in M}, \{ X(s), Q(\hat{s}', s) \}_{s \in S} \}$ all observed agent characteristics. Then the likelihood for observation $i$ is given by

$$\int_{\Theta} \left[ \prod_{j \in M} f(M(j) \mid D, \theta; \psi) \prod_{s \in S} \left\{ f(Y(s) \mid D, \theta; \psi) \Pr(G(s) = 1 \mid D, \theta; \psi) \right\} \right]^{1 \{ s \in \Gamma \}} dF(\theta), \quad (10)$$

where $\Theta$ is the support of $\theta$. After taking the logarithm of equation (10) and summing across all agents, we obtain the sample log likelihood.

Let $\phi_{\sigma}(\cdot)$ denote the probability density function and $\Phi_{\sigma}(\cdot)$ the cumulative distribution function of a normal distribution with mean zero and variance $\sigma$. The density functions for measurement and earning equations take a standard form conditional on the factors and other relevant observables:

$$f(M(j) \mid \theta, X(j)) = \phi_{\sigma_{j}(\cdot)}(M(j) - X(j)'\kappa_j - \theta'\gamma_j) \quad \forall \ j \in M$$
$$f(Y(s) \mid \theta, X(s)) = \phi_{\sigma_{s}(\cdot)}(Y(s) - X(s)'\beta_s - \theta'\alpha_s) \quad \forall \ s \in S.$$

The derivation of the transition probabilities has to account for forward-looking agents who make their educational choices based on the current costs and expectations of future rewards. Agents know the full cost of the next transition and the systematic parts of all future outcomes and costs $(X(s)'\beta_s, Q(\hat{s}', s)'\delta_{\hat{s}', s})$. They do not know the values of future random shocks. Agents at state $s$ decide whether to transition to the costly state $\hat{s}'$ or the no-cost alternative $\tilde{s}'$. Their ex ante valuations $T(s')$, $s' \in (\hat{s}', \tilde{s}')$, incorporate expected earnings and costs, and the continuation value $CV(s')$ from future opportunities. Given our functional form assumptions, the ex ante value of state $s'$ is:

$$T(s') = \begin{cases} 
X'(\hat{s}')\beta_{\hat{s}'} + \theta'\alpha_{\hat{s}'} - Q(\hat{s}', s)'\delta_{\hat{s}', s} - \theta'q_{\hat{s}', s} + CV(\hat{s}') & \text{if } s' = \hat{s}' \\
X'(\tilde{s}')\beta_{\tilde{s}'} + \theta'\alpha_{\tilde{s}'} + CV(\tilde{s}') & \text{if } s' = \tilde{s}'.
\end{cases}$$
The ex ante state evaluations and distributional assumptions characterize the transition probabilities:

\[
\Pr\left( G(s') = 1 \mid D, \theta; \psi \right) = \begin{cases} 
\Phi_{\psi(s', s)} \left( T(s') - T(\tilde{s}') \right) & \text{if } s' = \tilde{s}' \\
1 - \Phi_{\psi(s', s)} \left( T(s') - T(\tilde{s}') \right) & \text{if } s' = \hat{s}' .
\end{cases}
\]

Finally, the continuation value of \( s \) is:

\[
CV(s) = \left[ \Phi_{\psi(s', s)} \left( T(s') - T(\tilde{s}') \right) \right] \times 
\int_{-\infty}^{T(\hat{s}') - T(\tilde{s}')} 
\left[ T(\hat{s}') - \eta \right] \Phi_{\psi(s', s)}(\eta) \Phi_{\psi(s', s)}(T(s') - T(\tilde{s}')) d\eta
+ \left[ 1 - \Phi_{\psi(s', s)} \left( T(s') - T(\tilde{s}') \right) \right] \times T(\hat{s}'),
\]

where we integrate over the conditional distribution of \( \eta(s', s) \) as the agent chooses the costly transition to \( \hat{s}' \) only if \( T(s') - \eta(s', s) > T(\tilde{s}') \).

We compare ML against SMM for statistical and numerical reasons. ML estimation is fully efficient as it achieves the Cramér-Rao lower bound. In addition, numerical integration, evaluation of the normal cumulative distribution function, and all arithmetic calculations are of high precision in our ML implementation. The program is written in long double precision. We require high quality numerical integration as we integrate \( \theta \) out of the likelihood function. The accuracy of alternative integration strategies depends on the point of evaluation, so we check for the most appropriate quadrature rule at each step of the maximization. Additional discussion on both issues is presented in Web Appendix E. For the evaluation of the normal cumulative distribution function we use the algorithm proposed by Marsaglia (2004). The program is accurate up to 15 digits. We provide our analysis of the precision in Web Appendix A. We maximize the sample log likelihood using the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm (Press et al., 1992). Our program allows for the parallel evaluation of agent likelihood contributions.

4.2 The SMM Approach

We present the basic idea of the SMM approach and the details of the criterion function. Then we discuss the choice of tuning parameters. The goal in the SMM approach is to choose a set of structural parameters \( \psi \) to minimize the weighted distance between selected moments from the observed sample.
and a sample simulated from a structural model. The criterion function takes the following form:

$$\Lambda(\psi) = \left[ \hat{f} - \hat{f}(\psi) \right]' W^{-1} \left[ \hat{f} - \hat{f}(\psi) \right],$$

(11)

where $\hat{f}$ represents a vector of moments computed on the observed data and $\hat{f}(\psi)$ denotes an average vector of moments calculated from $R$ simulated datasets and $W$ is a positive definite weighting matrix. We define $\hat{f}(\psi)$ as:

$$\hat{f}(\psi) = \frac{1}{R} \sum_{r=1}^{R} \hat{f}_r(u_r; \psi).$$

The simulation of the model involves the repeated sampling of the unobserved components $u_r = \{ \{ \epsilon(s), \eta(s', s) \} \}_{s \in S}$ determining agents’ outcomes and choices. We repeat the simulation $R$ times for fixed $\psi$ to obtain an average vector of moments. $\hat{f}_r(u_r; \psi)$ is the set of moments from a single simulated sample. We solve the model through backward induction and simulate 5,000 educational careers to compute each single set of moments. We keep the conditioning on exogenous agent characteristics implicit in equation (11).

We account for the presence of $\theta$ by estimating a vector of factor scores based on $M$ that proxy the latent skills for each participant (Bartlett, 1937). The scores are subsequently treated as ordinary regressors in the estimation of the auxiliary models. We use the true factors in the simulation steps, assuring that SMM and ML are correctly specified.

The random components $u_r$ are drawn at the beginning of the estimation procedure and remain fixed throughout. This avoids chatter in the simulation for alternative $\psi$, where changes in the criterion function could be due to either $\psi$ or $u_r$ (McFadden, 1989).

To implement our criterion function it is necessary to choose a set of moments, the number of replications, a weighting matrix, and an optimization algorithm. Later, we investigate the sensitivity of our results to these choices. We use 250 moments to estimate a total of 138 free structural parameters. We classify the moments into two groups. The first group captures within state returns to schooling, in-

---

cluding the mean and standard deviation of within state wages and the parameters of Ordinary Least Squares (OLS) regressions of wages on covariates. The second group relates to educational choices and their costs, including the state frequencies and the parameters of a Linear Probability (LP) model of educational choices on cost shifters $Q(\hat{s}',s)$ and future outcome covariates $X(s')$ for each transition. We include values of $X(s')$ in the set of regressors to approximate the dynamic nature of agents’ choices. We set the number of replications $R$ to 50 and thus simulate a total of 250,000 educational careers for each evaluation of the criterion function. The weighting matrix $W$ is a diagonal matrix with the variances of the moments. We determine the latter by resampling the observed data 200 times. We exploit that our criterion function has the form of a standard nonlinear least-squares problem in our optimization. Due to our choice of the weighting matrix, we can rewrite equation (11) as

$$\Lambda(\psi) = \sum_{i=1}^{I} \left( \frac{f_i - \hat{f}_i(\psi)}{\hat{\sigma}_i} \right)^2,$$

where $f_i$ denotes moment $i$ and $\hat{\sigma}_i$ its bootstrapped standard deviation.

Our criterion function is not a smooth function of the model parameters. Small changes in the structural parameters cause some simulated agents to change their educational choices, resulting in discrete jumps in our set of moments (Keane and Smith, 2003). Thus we cannot use gradient-based methods for optimization and rely on derivative-free alternatives instead. Moré and Wild (2009) show that model-based solvers perform better than standard derivative-free direct search solvers used in the existing literature (Adda et al., 2013; 2011; Del Boca et al., 2013). From the class of model-based solvers, we choose the Practical Optimization Using No Derivatives for Sums of Squares (POUNDerS) algorithm (Munson et al., 2012). POUNDerS exploits the special structure of the nonlinear least-squares problem within a derivative-free trust-region framework and forms a smooth approximation model of the objective function to converge to a minimum.\textsuperscript{12} Our program allows for the parallel simulation of agents within each evaluation of the criterion function.

\textsuperscript{12}See Nocedal and Wright (2006) for a discussion of the nonlinear least-squares problem and Kortelainen et al. (2010) for a detailed description of the underlying mechanics of POUNDerS.
4.3 Results

We compare ML and SMM estimation to learn whether our version of SMM is a good substitute for ML. First, we compare basic model fit statistics. Second, we study the estimates for the gross and net returns to education. Finally, we explore alternative choices for the set of moments, weighting matrix, number of replications, and optimization algorithm.

Model Fit Table 4 shows the average annual earnings for each state and the conditional state frequencies. Overall, both estimation approaches fit these aggregate statistics quite well. The model fit for the average earnings among late college graduates and late college dropouts is slightly worse than for the other states as the agent count in those states is low. This affects the SMM estimates more than ML. The state frequencies are identical between the observed and simulated datasets.

We report the root-mean-square error (RMSE) based on the difference between the simulated and observed statistics. There are only minor discrepancies for both estimation approaches. Nevertheless, they are slightly smaller for the ML results.
Table 4: Cross Section Model Fit

<table>
<thead>
<tr>
<th>State</th>
<th>Average Earnings</th>
<th>State Frequencies</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Observed</td>
<td>ML</td>
</tr>
<tr>
<td>High School Dropout</td>
<td>2.50</td>
<td>2.52</td>
</tr>
<tr>
<td>High School Finishing</td>
<td>0.77</td>
<td>0.77</td>
</tr>
<tr>
<td>High School Graduation</td>
<td>2.46</td>
<td>2.45</td>
</tr>
<tr>
<td>Early College Enrollment</td>
<td>1.45</td>
<td>1.43</td>
</tr>
<tr>
<td>High School Graduation (cont’d)</td>
<td>3.87</td>
<td>3.89</td>
</tr>
<tr>
<td>Late College Enrollment</td>
<td>2.51</td>
<td>2.55</td>
</tr>
<tr>
<td>Early College Graduation</td>
<td>6.80</td>
<td>6.76</td>
</tr>
<tr>
<td>Early College Dropout</td>
<td>3.90</td>
<td>3.94</td>
</tr>
<tr>
<td>Late College Graduation</td>
<td>6.03</td>
<td>6.18</td>
</tr>
<tr>
<td>Late College Dropout</td>
<td>5.10</td>
<td>5.04</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>ML</th>
<th>SMM</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMSE</td>
<td>0.0332</td>
<td>0.05094</td>
</tr>
</tbody>
</table>

Notes: Earnings are discounted using the within state duration and measured in units of $10,000. Statistics calculated for ML and SMM approaches based on 50,000 simulated agents using the parameter estimates. RMSE = root-mean-square error.

We apply $\chi^2$ goodness of fit tests (Heckman, 1984; Heckman and Walker, 1990) to the estimated and actual probabilities. In Table 5 we report the $p$-value of a joint test of the relative share of agents within each state conditional on all possible realizations of selected covariates.\footnote{In the $\chi^2$ test, the predicted conditional distributions depends on estimated parameters. We do not adjust the test statistic to account for parameter estimation error as suggested by (Heckman, 1984) because the adjustments are usually slight (Heckman and Walker, 1990).}
Overall, the level of \( p \)-values is high. For ML estimation, all \( p \)-values indicate that our model is consistent with the data at the 5% significance level. In the case of SMM, we only do not pass the test conditional on Number of Children among early college enrollees. Because tests within covariates across all states are not independent, we use a Bonferroni test to evaluate the joint hypothesis that the predicted covariate distributions fit at each state. The test is based on the maximum \( \chi^2 \) statistic over all states for each covariate. We pass a 5% Bonferroni test for all covariates and both estimation approaches.

### Economic Implications
Table 6 presents the median *ex ante* gross returns \( GR^a(\hat{s}', s', s) \) and net returns \( NR^a(\hat{s}', s', s) \) of pursuing a higher education by transitioning from \( s \) to \( \hat{s}' \). Both capture all current and future earnings. However, they differ with regards to current and future costs. Their systematic parts are included in the calculation of the \( NR^a(\hat{s}', s', s) \) but not the \( GR^a(\hat{s}', s', s) \) as we discussed in Section 2.2.
Table 6: Economic Implications

<table>
<thead>
<tr>
<th>State</th>
<th>Gross Return</th>
<th>Net Return</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>True ML SMM</td>
<td>True ML SMM</td>
</tr>
<tr>
<td>High School Finishing</td>
<td>28% 35% 37%</td>
<td>66% 62% 197%</td>
</tr>
<tr>
<td>Early College Enrollment</td>
<td>17% 17% 15%</td>
<td>-2% -2% -6%</td>
</tr>
<tr>
<td>Early College Graduation</td>
<td>71% 73% 82%</td>
<td>48% 46% 107%</td>
</tr>
<tr>
<td>Late College Enrollment</td>
<td>28% 30% 30%</td>
<td>-23% -21% -44%</td>
</tr>
<tr>
<td>Late College Graduation</td>
<td>22% 25% 28%</td>
<td>9% 8% 35%</td>
</tr>
</tbody>
</table>

ML SMM

RMSE

<table>
<thead>
<tr>
<th></th>
<th>ML</th>
<th>SMM</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMSE</td>
<td>0.07854</td>
<td>0.54404</td>
</tr>
</tbody>
</table>

Notes: Statistics calculated for ML and SMM approaches based on 50,000 simulated agents using the parameter estimates. RMSE = root-mean-square error calculated in units of 100%.

The estimates for the gross returns $GR^a(\hat{s}', \tilde{s}', s)$ are very similar for the two approaches and close to their true values. However, for the net returns $NR^a(\hat{s}', \tilde{s}', s)$ only the ML results are close to the truth. The SMM results are off by up to a factor of three. For example, the true net return of finishing high school is 66%, while SMM estimates 197%. The RMSE is roughly one order of magnitude larger for SMM than ML estimation. This difference is solely driven by the discrepancies in the net returns.

Table 7 sheds light on the poor performance of our SMM approach in the estimation of the net returns. These, in contrast to the gross returns, include the current costs and the systematic part of all future costs of educational choices. SMM is unable to detect the systematic differences in the cost faced by agents. We overestimate the variance of the unobserved component determining choices $\sigma_{\eta(\hat{s}', s)}$. Too much of the agents’ decisions is attributed to random cost shocks and not their systematic differences. This translates into an excess net return as we underestimate the cost associated with future educational choices. Despite encouraging values for model fit criteria, SMM fails to accurately estimate the net return to educational choices.
Table 7: Standard Deviations

<table>
<thead>
<tr>
<th>State</th>
<th>True</th>
<th>ML</th>
<th>SMM</th>
</tr>
</thead>
<tbody>
<tr>
<td>High School Finishing</td>
<td>0.27</td>
<td>0.24</td>
<td>0.85</td>
</tr>
<tr>
<td>Early College Enrollment</td>
<td>0.20</td>
<td>0.18</td>
<td>0.78</td>
</tr>
<tr>
<td>Early College Graduation</td>
<td>0.61</td>
<td>0.59</td>
<td>1.50</td>
</tr>
<tr>
<td>Late College Enrollment</td>
<td>0.22</td>
<td>0.20</td>
<td>0.47</td>
</tr>
<tr>
<td>Late College Graduation</td>
<td>0.61</td>
<td>0.59</td>
<td>1.50</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.01605</td>
<td>0.68134</td>
<td></td>
</tr>
</tbody>
</table>

Notes: RMSE = root-mean-square error.

We now investigate the poor performance of our application of SMM and start with some evidence that we indeed recover a global minimum of our criterion function. Figure 7 shows the value of \( \Lambda(\hat{\psi}) \) around our SMM estimates as we perturb all parameters in a random direction in \( t \) increments. All perturbations increase the discrepancies between the observed and simulated sample. However, \( \Lambda(\hat{\psi}) \) is not zero because of remaining differences between estimated and true structural parameters. Even if we set \( \hat{\psi} = \psi^* \), then \( \Lambda(\psi^*) \) evaluates at 232 (horizontal dashed line) due to the random variation in agents’ behaviors and state experiences. The moments provide noisy information about the data generating process due to the random components. The more unobserved variation in the data, the less information they contain. This is why our criterion function does not attain its global minimum when evaluated at \( \psi^* \) but at \( \hat{\psi} \). Next we consider alternative choices for: (1) set of moments \( \hat{f}(\psi) \), (2) number of replications \( R \), (3) weighting matrix \( W \), and (4) optimization algorithm.

Set of Moments We followed the common practice of selecting our set of moments in an ad hoc fashion while appealing to basic relationships between moments and structural parameters (Adda et al., 2013, 2011; Flinn and Del Boca, 2012). In our baseline, we include aggregate statistics of the data such as average earnings and their standard deviations as well as state frequencies. We also match
a number of conditional moments such as parameters of OLS regressions for within state earnings
and LP models characterizing the state transitions. We explicitly included future values $X(s')$ among
the regressors in the LP models to capture the dynamics of agents’ educational choices. In Table 8,
we study alternative sets of moment conditions. In particular, we specify a cross sectional version
in which we do not include future outcome covariates in the models of educational choice. We also
study three alternative sets of dynamic moments. We increase their number from 250 up to 545, adding
moments that provide additional information about the observed agent transitions. We thereby hope
to improve the estimation of the systematic differences in the psychic cost of educational choices. We
add a dynamic Probit model for each transition (Alt. A) and correlations of within states outcomes
and each covariate $(Y(s), X(s))$, between outcomes over time $(Y(s), Y(s'))$, and correlations of choice
indicators with each current cost and future earnings covariate $(G(s), (Q(s', s), X(s')))$ (Alt. B).
Table 8: Set of Moments

<table>
<thead>
<tr>
<th>Sets</th>
<th>Cross Section Moments</th>
<th>Dynamic (Panel) Moments</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Base</td>
<td>Base</td>
</tr>
<tr>
<td>Outcome Models</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Means</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Standard Deviations</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Ordinary Least Squares</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Correlations</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Choice Models</td>
<td></td>
<td></td>
</tr>
<tr>
<td>State Frequencies</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Linear Probability</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- cross section</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>- dynamic</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Probit</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- dynamic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correlations</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall Statistics</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of Moments</td>
<td>222</td>
<td>250</td>
</tr>
<tr>
<td>Number of Replications</td>
<td></td>
<td>50</td>
</tr>
<tr>
<td>Weighting Matrix</td>
<td>diagonal variance matrix</td>
<td></td>
</tr>
<tr>
<td>Algorithm</td>
<td>POUNDerS</td>
<td></td>
</tr>
<tr>
<td>Quality of Fit Measures</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\Lambda(\hat{\psi})$</td>
<td>632.07</td>
<td>218.32</td>
</tr>
<tr>
<td>$\Lambda(\psi^*)$</td>
<td>215.12</td>
<td>232.66</td>
</tr>
</tbody>
</table>

Notes: Alt. = Alternative.
We also report the value of the criterion function at the true structural parameters $\Lambda(\psi^*)$. Its difference from zero is solely driven by the presence of the random disturbances $u_r$. The final values of our criterion function, at least when using dynamic moments, are always below $\Lambda(\hat{\psi}^*)$ which gives us further confidence that we attained a global minimum in those cases. However, when using cross sectional moments only, our optimizer stops above $\Lambda(\hat{\psi}^*)$.

We show the implications of alternative moments for the estimated median \textit{ex ante} gross and net returns to education in Table 9. The configuration of all alternative sets of moments is presented in Table 8.

### Table 9: Robustness of Economic Implications of Alternative Implementations of SMM

<table>
<thead>
<tr>
<th>State</th>
<th>Cross Section Moments</th>
<th>Dynamic (Panel) Moments</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>True</td>
<td>Base</td>
</tr>
<tr>
<td><strong>Gross Return</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High School Finishing</td>
<td>28%</td>
<td>16%</td>
</tr>
<tr>
<td>Early College Enrollment</td>
<td>17%</td>
<td>-6%</td>
</tr>
<tr>
<td>Early College Graduation</td>
<td>71%</td>
<td>57%</td>
</tr>
<tr>
<td>Late College Enrollment</td>
<td>28%</td>
<td>41%</td>
</tr>
<tr>
<td>Late College Graduation</td>
<td>22%</td>
<td>16%</td>
</tr>
<tr>
<td><strong>Net Return</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High School Finishing</td>
<td>66%</td>
<td>215%</td>
</tr>
<tr>
<td>Early College Enrollment</td>
<td>-2%</td>
<td>-3%</td>
</tr>
<tr>
<td>Early College Graduation</td>
<td>48%</td>
<td>79%</td>
</tr>
<tr>
<td>Late College Enrollment</td>
<td>-23%</td>
<td>-84%</td>
</tr>
<tr>
<td>Late College Graduation</td>
<td>9%</td>
<td>26%</td>
</tr>
</tbody>
</table>

\textbf{Notes:} Statistics calculated for SMM based on 50,000 simulated agents using the parameter estimates.
Once dynamic moments are included in the criterion function, the effect of adding even more is rather small. The estimates for the gross and net returns are all very similar. However, when using only cross sectional moments for the criterion function, the performance of SMM deteriorates further and its ability to recover gross returns is undermined. The difference between the estimated and true gross returns is always larger (in absolute value) when using cross sectional moments only. There is no such clear pattern for the net returns, which are poorly estimated in all specifications.

We assess the information content of selected moments \( \hat{f}_i \) and investigate the effect of perturbations (in percentage terms) around \( \hat{\psi} \). As examples we consider changes in structural intercepts for outcomes and costs. Figure 8 shows the effect of changes in the structural cost intercept of finishing high school on the probability that agents complete their high school education (Figure 8a) and the intercepts of a dynamic LP (Figure 8b) and Probit (Figure 8c) model for that transition. Given our baseline SMM estimate, 86% of agents finish high school while 14% drop out. A 50% increase in the structural cost intercept decreases the number of finishers to 81%, while a 50% decrease boosts it to 90%. These trends are also captured by the changes in the intercept of the dynamic LP and Probit model. It is clear from a comparison of Figure 8b and 8c why the addition of the Probit model does not improve the performance of the estimation. There is no new information added to the set of moments because the perturbations of structural parameters translate into similar changes in the moments.

In Figure 9, we perturb the intercept in the structural wage equation for early college graduates. This has a direct effect on average wages in that state (Figure 9a). However, agents are forward-looking and these changes also affect moments associated with earlier decisions such as finishing high school (Figure 9b). This is true even though the immediate benefits of doing so (Figure 9c) are unaffected. Agents change their early educational choices due to the increase in the option value of finishing high school, which includes the expected future value of potentially graduating from college.

**Number of Replications** For a given set of structural parameters, we create multiple simulated datasets from which we calculate the moments. Averaging over those moments, we reduce the effect of random components determining agents’ choices and state experiences. In Figure 10 we show the value of the criterion function at \( \phi^* \) for different numbers of replications \( R \). The difference from zero is solely driven by the random components determining agents’ choices and outcomes. If the model
is simulated only once, then $\Lambda(\psi^*)$ takes value 465. Initially, increases in $R$ result in a large drop of $\Lambda(\psi^*)$. However, this effect levels off after more than 40 replications. Afterwards, the value of $\Lambda(\psi^*)$ oscillates around 230. In a finite sample, differences between $\hat{f}$ and $\tilde{f}(\psi^*)$ remain even for a very large number of replications. While the random values of $(\epsilon(s), \eta(s', s))$ wash out in the simulated moments, their particular realizations remain relevant in the finite observed data. For our baseline estimates we set $R = 50$. Further increases do not change model fit or economic implications.

**Weighting Matrix** Our optimization algorithm is only guaranteed to converge to local minimizers. Figure 11 plots the surface of our criterion function around $\psi^*$ for two alternative choices of $W$ given the true values of $u_T$. Thus, $\hat{f} = \hat{f}(\psi^*)$ and $\Lambda(\psi^*)$ evaluates initially to zero regardless of the weighting matrix used. Then we perturb all the structural parameters in a random direction in $t$ increments. We
Figure 9: Parameter Perturbations, Outcome

(a) Early College Graduation, Average Wages

(b) High School Finishing, State Frequency

(c) High School Finishing, Average Wages

Figure 10: Role of Replications

Notes: Investigation using estimation sample of 5,000 agents with varying number of replications.
show the surface of $\Lambda(\psi)$ when either the identity matrix (Figure 11a) or the diagonal matrix with the variances of the moments (Figure 11b) is used. Choosing the identity matrix for $W$ results in multiple local minima, whereas using the variances smoothes the overall surface of the criterion function.

**Figure 11: Alternative Weighting Matrices**

![Figure 11a: Identity Matrix](image1)
![Figure 11b: Inverse Variances on Diagonal](image2)

**Notes:** Investigation using estimation sample of 5,000 agents with 50 replications and alternative weighting matrices.

**Optimization Algorithm** Because we repeat the SMM estimation many times for our Monte Carlo study, we benefit from a faster optimization algorithm. In Figure 12 we compare the performance of POUNDerS to the standard Nelder-Mead algorithm (Nelder and Mead, 1965) applied by Del Boca et al. (2013) and French and Jones (2011) among others. We perturb our estimates $\hat{\psi}$ and run the two algorithms as implemented in the Toolkit for Advanced Optimization (TAO) (Munson et al., 2012) to investigate their relative performance. Following More and Wild (2009) we use the default input and algorithmic parameters. Both algorithms are derivative-free, but differ in their search strategy and how they exploit the structure of the criterion function. Nelder-Mead applies a direct search method, while POUNDerS forms an approximation model within a trust region which exploits the special structure of our nonlinear least-squares problem. We show a minute-by-minute account of the criterion function $\Lambda(\psi)$ over five hours.

---

14We are aware that performance can change for other choices. However, our practical experience throughout this project lines up with the results from this stylized presentation. We illustrate the relative performance of the two algorithms using a single processor only. Both algorithms allow parallel implementations as well. (Lee and Wiswall, 2007; Munson et al., 2012).
The POUNDerS algorithm attains a lower bound of $\Lambda(\psi) \approx 300$ after two hours. With the Nelder-Mead algorithm, the criterion function still takes a value of $\Lambda(\psi) \approx 2,600$ after five hours. Even after 36 hours, the Nelder-Mead solution $\Lambda(\psi) \approx 1,126$ is still about four times as large as the POUNDerS solution.

We are unable to improve the SMM results by using alternative tuning parameters. Our discussion cautions that inspection of model fit statistics alone does not guarantee accurate economic implications. For our model, the Monte Carlo exercise reveals that large unobserved variation in educational choices leaves SMM unable to recover the true returns to education. The structural variances of the unobserved cost shocks are poorly estimated. In ongoing research we are exploring two strategies to improve this weakness of our SMM implementation. First, we are experimenting with alternative moments that provide additional unique information about the unobservable cost shocks. Second, we are examining the usefulness of stochastic optimization (Spall 2003) for simulation-based estimation of structural models. This approach explicitly accounts for random objective functions.
5 Conclusion

We compare the performance of simulated method of moments (SMM) and maximum likelihood (ML) estimation in dynamic discrete choice models. We formulate a deliberately simplistic dynamic model of educational choices which emphasizes the role of unobserved heterogeneity, psychic costs, and option values for the net returns to schooling. The primary value of the model is as input to the simulation study that is the core of this paper.

We estimate our model on a sample of white males from the National Longitudinal Survey of Youth of 1979 (NLSY79). We discuss its implications for schooling decisions and present estimates of option values by cognitive and non-cognitive factors. Given our estimates, we simulate a synthetic sample, creating a realistic setting to compare ML and SMM estimation. Our model allows for ML estimation without the need for any simulation of the likelihood function, which provides a clean comparison of ML against simulation-based estimation methods such as SMM. We analytically characterize the accuracy of the ML estimation algorithm and establish that it is accurate to at least three digits of precision in estimating parameters. ML and SMM pass standard model fit tests. However, while the ML estimates are close to the true structural objects of interest, our version of SMM fails to recover the true net returns to education. The SMM substantially underestimates psychic costs.

We investigate alternative tuning parameters for our SMM procedure. We specify alternative sets of moment conditions and show how the benefit of additional moments depends on the unique information they provide. Moments that capture the dynamics of agent behavior are crucial for getting reliable estimates of dynamic models. A large replication count in the simulation step reduces the effect of random noise in the measurement of the criterion function. An appropriate choice of the weighting matrix smoothes the surface of the criterion function and reduces the risk of local minima. Our choice of a fast state of the art optimization algorithm allows for a variety of additional robustness checks even in our computation intensive model. Based on our analysis we recommend that more exacting model specification tests be used to verify the robustness of the performance of SMM in any application.
References


