ELSEVIER

Journal of Econometrics 96 (2000) 201–229

JOURNAL OF Econometrics

www.elsevier.nl/locate/econbase

# Semiparametric identification and heterogeneity in discrete choice dynamic programming models

Christopher R. Taber

*Department of Economics and Institute for Policy Research Northwestern University, 2003 Sheridan Road, Evanston, IL 60208, USA*

Received 1 January 1997; received in revised form 1 June 1999

**Abstract**

Empirical discrete choice dynamic programming models have become important empirical tools. A question that arises in estimation and interpretation of the results from these specifications is which combination of data and assumptions are needed to overcome problems of heterogeneity, selection, and omited variables bias. This paper addresses this question by considering nonparametric identification of a version of the model that allows for quite general forms of unobservable and information structures. I show that the model can be identified under conditions similar to a static polychotomous choice model. Using a stochastic version of an 'identification of infinity' argument, utility can be identified up to a monotonic transformation of the observables under strong support conditions and two types of exclusion restriction. The first type is similar to a standard static exclusion restriction: a variable that influences the first period decision, but does not enter the second period decision directly. The second type requires a variable that does not affect the utility of the first option directly, but is known during the first period, and has predictive power on the choice during the second. I also provide two specifications under which the full error structure can be identified. This requires the additional assumption of stochastic innovations in the observables. I then use the model to estimate schooling decisions in which students deciding whether to drop out of high school account for the option value of attending college. 

*JEL classification:* C14; C35; C51

*Keywords:* Identification; Discrete choice; Dynamic programming


PII: S0304-4076(99)00057-3

## 1. Introduction

Empirical discrete choice dynamic programming models have become important empirical tools. In some applications of these models, problems of substantial heterogeneity/selection/omitted variable bias arise (see, e.g. Keane and Wolpin (1997) or Eckstein and Wolpin (1997)). The source of these biases is potentially more complex in dynamic models than static ones in that agents may have heterogeneity not only in outcomes, but also in expectations about future outcomes. A question that arises in estimation and interpretation of the results in these cases is which combination of data and assumptions are needed to overcome these problems. This paper addresses this question by considering nonparametric identification of a version of the model that allows for quite general forms of unobservable and information structures. Despite the added complexity of the model, I show that it can be identified under conditions similar to a static polychotomous choice model. Using a stochastic version of an 'identification of infinity' argument, utility can be nonparametrically identified up to a monotonic transformation of the observables under strong support conditions and two types of exclusion restriction. The first type is similar to a standard static exclusion restriction: a variable that influences the first period decision, but does not enter the second period decision directly. The second type requires a variable that does not affect the utility of the first option directly, but is known during the first period and has predictive power on the choice during the second. I also provide two specifications under which the full error structure can be identified. This requires the additional assumption of stochastic innovations in the $X$'s: a variable known at time one that helps predict the second period decision, but conditional on second period observables, has no influence on the decision.

The specification I develop is a generalization of a dynamic 'Roy' type model, and I focus on schooling decisions. In deciding whether to drop out of high school a student takes into account both the direct value of graduating from high school as well as the value of the option to attend college. While making this decision, a student does not know whether he will attend college. Heterogeneity bias is likely to be important in that students with high returns or tastes for high school are also likely to have high returns or tastes for college. While there is a substantial literature addressing the selection/heterogeneity issue in schooling models, the previous work has typically ignored the complexity of the heterogeneity. The problem is not just that the returns to college are likely to be correlated with returns to high school, but also that agents may have additional information about their own private returns to college which is unobservable to the econometrician. For example, a high school student may know that he has excellent teaching skills. While this information may be correlated with the returns to high school, since teachers must have a college degree it is much more important for the decision about whether to attend college. Accounting for this type of heterogeneity in information requires a more complex information structure about unobservables than is often used in empirical work. This leads to two important questions (1) can an information structure such as this be identified? and (2) if not, can other important structural parameters be identified without this information? I provide a set of conditions under which the coefficients can be identified allowing for these forms of unobserved heterogeneity in information about the unobservables. While we can not identify an arbitrarily complicated information structure under standard conditions, I provide two specifications under which we can.

Discrete choice dynamic programming models have been applied to a large range of topics. Examples include patent renewal (Pakes, 1986), bus engine replacement (Rust, 1987), job search (Wolpin, 1987), fertility (Hotz and Miller, 1993), life cycle earnings (Keane and Wolpin, 1997), and schooling (Taber, 1998); a survey can be found in Eckstein and Wolpin (1989) or Rust (1994). The main goal of this paper is to establish identification of these models under fairly weak assumptions about the distribution of the error and information structure. These results are useful for two reasons. First, they take a first step towards semiparametric estimation of this class of models by establishing sufficient conditions for their identification. To facilitate estimation, this work typically imposes strong parametric restrictions on the distribution of the unobservables and on the information structure that agents use to form their expectations. These assumptions are typically chosen out of mathematical convenience rather than as implications of the models themselves so it is important to check the sensitivity of the model to these assumptions. Secondly, and perhaps more importantly given current computational problems, they demonstrate the ideal data set under which these models can be identified without parametric restrictions. Solving the heterogeneity bias problem can typically be achieved by imposing functional forms on the distribution of the error terms. However, it is preferable to find data that can solve the problems. In practice the perfect data set rarely exists, so identification is achieved through a combination of data and assumptions. Nevertheless, this type of identification exercise is potentially useful both for understanding the trade-off between assumptions and data and for illuminating which type of data one should use when estimating these models.

While much work has been done on semiparametric identification of other discrete choice models, it has not been systematically discussed in dynamic programming problems. There have been a few papers that focus on specific points, often with negative results. Flinn and Heckman (1982) consider identification of job search models. They show that these models are nonparametrically underidentified as one essentially can not distinguish high reservation wages from low arrival rates. Rust (1994) also shows a form of non-identification in a more general model. As I discuss below, this problem can be addressed fairly easily in a finite time model, but is a more serious concern in infinite time models. The most closely related work is by Pakes and Simpson (1989). They provide a sketch of identification for a finite period model of patent renewal that could be written as a special case of mine. They also use exclusion restrictions and essentially a similar identification at infinite argument.[1] I extend this model into a broader framework by allowing for a more general form for unobservables and information, and a more general process for the observables. Cameron and Heckman (1998) also consider identification of schooling models, but the form of their models are quite different in that they do not use this dynamic programming framework.

This paper extends the work on identification of discrete selection models in static cases to incorporate dynamics. As in this paper, most of the previous work generalizes the ideas behind the semiparametric identification of the binary choice model,

$$d = 1(g(X, \theta) + \varepsilon > 0). \tag{1}$$

The function $g$ is assumed known up to parameter $\theta$, but the distribution of $\varepsilon$ is unspecified. Identification of this simple model is presented in Cosslett (1983) and Manski (1975,1988). Extensions that allow for multiple choices or multiple periods include Manski (1987), Thompson (1989), Cameron and Heckman (1998), and Cameron and Taber (1994). Matzkin (1990,1992,1993) follows another line. She extends the semiparametric identification to nonparametric identification. For instance in the binary choice model (1) she allows $g(X, \theta) = g(X)$ and provides conditions under which the function $g$ is identified.

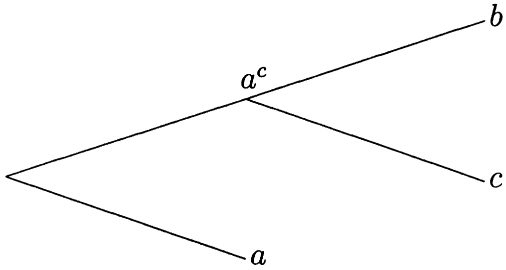
I describe the model in Section 2. I provide identification of various components of the model in Sections 3 and 4. In Section 5 I demonstrate how these results can be used by estimating a version of the model as a schooling model where a student first decides whether to graduate from high school and then conditional on high school graduation decides whether to attend college. Section 6 presents some conclusions.

## 2. Model and notation

In order to concentrate on the issues arising from unobserved heterogeneity, I use a dynamic programming specification that is as simple as possible in terms of the numbers of periods and choices, but quite general in terms of the joint distribution of unobservables and information possessed by the agents. Extending the results below to more complex finite horizon specifications is straight

[1] They do not explicitly talk about identification at infinity, but in their notation send a second period variable $c_2$ above $Y_J$ where $Y_J$ is the upper bound of the support of the second period error term.

forward. The model consists of two time periods and three terminal states. It can be thought of essentially as a dynamic extension of the Roy model (Roy, 1951; Heckman and Honoré, 1990). The structure takes the following form,




In the first period the agent chooses between node $a$ and node $a^c$. If she chooses node $a^c$ in the first period, she then chooses between nodes $b$ and $c$ in the second. In a schooling model node $a$ could represent dropping out of high school, node $b$ graduating from high school and entering the labor force, and node $c$ graduating from college. When making the decision to graduate from high school the student does not know her college options with perfect certainty.

The agent's preferences are summarized by lifetime reward function $V_k$ at each terminal state $k \in \{a, b, c\}$. By defining utility at the terminal nodes, I do not separate utility at node $a^c$ from utility at nodes $b$ and $c$. Rust (1994) essentially shows that one can never distinguish utility incurred at node $a^c$ from utility incurred at node $b$ and $c$, but known with perfect certainty at time 1. The intuition behind this is clear in the schooling example in which it is impossible to tell whether the utility accumulated from graduating from high school actually is realized during the graduation ceremony or whether it accrues later in life while looking at the degree on the wall.[2] In an infinite time model this type of normalization is not possible, so the potential problem is more severe.

I define $V_a$ so that it is known at the time the choice between $a$ and $a^c$ is made and $V_b$ and $V_c$ are known at the time the choice between $b$ and $c$ is made. Let $\mathscr{I}_1$ denote the information available to the agent at the time of the first decision. I assume that decisions are made in order to maximize expected lifetime reward. Thus the reward function at node $a^c$ in the first period takes the value,

$$V_{a^c}(\mathscr{I}_1) \equiv \mathrm{E}[\max\{V_b, V_c\} \mid \mathscr{I}_1].$$

The agent chooses node $a$ if $V_a > V_{a^c}(\mathscr{I}_1)$, and chooses node $a^c$ otherwise. If she chooses $a^c$ in the first period, she chooses node $b$ in the second if $V_b > V_c$.

The incomplete information structure distinguishes this dynamic programming specification from static discrete choice specifications. Under perfect

[2] This assumes that at the time of graduation the student recognizes she will receive this utility later in life.

certainty, the agent would simply choose the alternative with the highest lifetime value function and the model would be identical to the standard polychotomous choice problem. The basic structure of the specification I present below is similar to the polychotomous choice models of McFadden (1981), Thompson (1989), or Matzkin (1993). The model differs from these others only in that during the first period the agents are uncertain about their utilities in the second.

The econometrician observes $(d_a, d_b, d_c, X_a, X_b)$, where for $k = \{a, b, c\}$, $d_k$ is an indicator that state $k$ was chosen. I define them explicitly as,

$$d_a = 1(V_a > V_{a^c}(\mathscr{I}_1)), \tag{2}$$

$$d_b = 1(V_a \leqslant V_{a^c}(\mathscr{I}_1), V_b > V_c), \tag{3}$$

$$d_c = 1(V_a \leqslant V_{a^c}(\mathscr{I}_1), V_b \leqslant V_c), \tag{4}$$

where $1(\cdot)$ is the indicator function taking the value one if its argument is true and zero if it is false.

I define the reward functions at each terminal node to take the following form:

$$V_a = g_a(X_a) + \varepsilon_a, \tag{5}$$

$$V_b = g_b(X_b) + \varepsilon_b, \tag{6}$$

$$V_c = 0. \tag{7}$$

Since utilities are identified only up to monotonic transformations I normalize utility at node $c$ to zero (see Taber, 1996, for justification). The functions $g_a$ and $g_b$ may be finite dimensional as in Manski (1975) or Cosslett (1983) or infinite dimensional as in Matzkin (1990). The random vector $(X_a, X_b)$ is observed by the econometrician and independent of the unobserved random vector $(\varepsilon_a, \varepsilon_b)$.[3] The joint distribution of the error terms are left unspecified. I allow the information set $\mathscr{I}_1$ to be heterogeneous across individuals and do not restrict private information that is contained in $\mathscr{I}_1$ to be independent of the error terms.

To simplify the exposition I assume that the first period information about $\varepsilon_b$ can be completely summarized by a random variable $\varepsilon_1$ which is known to the agent during the first period, but not observed by the econometrician. The actual numeric value that $\varepsilon_1$ takes is irrelevant for this discussion, but in general

[3] An unappealing feature of this type of model is the separability and independence between the observables and unobservables. This assumption is chosen out of convenience, not as an implication of economic theory. Unfortunately, something analogous to it is necessary for identification. I use this specification for its simplicity and for direct comparison with previous work. The proofs in this paper can be easily altered to address other types of restrictions that have been used for the binary choice model.

we may want to think of it as a very large dimensional vector. All that is relevant about $\varepsilon_1$ is the information it provides. Similarly, the first period information about $X_b$ is completely contained in the observable random vector $X_1$.[4] Thus,

$$\mathscr{I}_1 = \sigma(\varepsilon_1, X_1),$$

where the notation $\sigma(Y)$ denotes the sigma algebra generated by a random variable $Y$. In addition, I assume that the observables $(X_1, X_a, X_b)$ are independent of the unobservables $(\varepsilon_1, \varepsilon_a, \varepsilon_b)$.

Since the agents know the value of $X_a$ during the first period,

$$\sigma(X_a) \subset \sigma(X_1).$$

I am not requiring that $X_a = X_1$, only that knowledge of $X_1$ is sufficient for knowledge of $X_a$.[5] Similarly I do not assume that $\varepsilon_a = \varepsilon_1$, only that

$$\sigma(\varepsilon_a) \subset \sigma(\varepsilon_1),$$

thus the agent may have private information about the values of future unobservables that is not contained in $\sigma(\varepsilon_a)$.

The structure is summarized in the following table:

| Known to the Agent at time one | Learned by the Agent at time two | Observed by the Econometrician |
|---|---|---|
| $\varepsilon_1$, $\varepsilon_a$ | $\varepsilon_b$ | $X_1, X_a$ |
| $X_1, X_a$ | $X_b$ | $X_b$ |
| $G(X_b \mid X_1)$ | | $G(X_b \mid X_1)$ |
| | | $d_a, d_b, d_c$ |
| Action taken: $d_a$ | Action taken: $d_b, d_c$ | |

Proving identification involves showing that the functions $(g_a, g_b)$ and aspects of the joint distribution of $(\varepsilon_1, \varepsilon_a, \varepsilon_b)$ are identified from the observed conditional probabilities. I also allow common shocks to occur between period one and period two. I denote the outcome of these shocks by $\Psi_2$. These could correspond to macro shocks which influence the outcomes of all individuals.

[4] By assuming that $X_1$ is observable, I do not allow the agents to be better at forecasting their future values of $X_b$ and $X_c$ than the econometrician. This assumption is not crucial for the results in Section 3, but would complicate identification of the full model in Section 4.

[5] In other words, I can construct some function $a$ such that $X_a = a(X_1)$, but in general I cannot find a function $a^{-1}$ such that $X_1 = a^{-1}(X_a)$.

The consequences of these types of shocks is to cause the ex-ante conditional distribution of $(\varepsilon_b \mid \varepsilon_1)$ which agents use to form $V_{a^c}$ to differ from the ex-post realization of $(\varepsilon_b \mid \varepsilon_1, \Psi_2)$. This could also represent departure from rational expectations. The econometrician observes $\Pr(a \mid X_1)$ and $\Pr(b \mid X_1, X_b, \Psi_2)$, defined as,

$$\Pr(d_a = 1 \mid X_1) = \Pr(g_a(X_a) + \varepsilon_a > V_{a^c}(X_1, \varepsilon_1) \mid X_1) \tag{8}$$

$$\Pr(d_b = 1 \mid X_1, X_b, \Psi_2) = \Pr(g_a(X_a) + \varepsilon_a \leqslant V_{a^c}(X_1, \varepsilon_1), g_b(X_b) + \varepsilon_b > 0 \mid X_1, X_b, \Psi_2). \tag{9}$$

The goal of this work is to provide conditions under which (8) and (9) are sufficient for identification of the functions $(g_a, g_b)$ and the joint distribution of the unobservables $(\varepsilon_1, \varepsilon_a, \varepsilon_b)$.

As mentioned above, there is an aspect to this problem that differentiates it from most previous work on both static and dynamic discrete choice models. Since $\varepsilon_1$ is person specific, the function $V_{a^c}(X_1, \varepsilon_1)$ is also person specific. This represents a different type of heterogeneity: heterogeneity across expectations as opposed to heterogeneity across outcomes. The only restriction imposed on information heterogeneity is that an agent's time one expectations of $\varepsilon_b$ are independent of the observables $X = (X_1, X_2)$.

## 3. Identification of $g_a$ and $g_b$ up to monotonic transformations

In this section I provide conditions which deliver identification of $g_a$ and $g_b$ with minimal assumptions about the distribution of the unobservables. I use a definition of identification that is analogous to Matzkin (1992). By identification of $g_a$ and $g_b$ up to a monotonic transformation, I mean that for any alternative functions and distribution of error terms $(g_a^*, g_b^*, \varepsilon_a^*, \varepsilon_1^*, \varepsilon_b^*)$ consistent with the observed probabilities,

$$\Pr[g_a(X_a) + \varepsilon_a > V_{a^c}(X_1, \varepsilon_1) \mid X] = \Pr[g_a^*(X_a) + \varepsilon_a^* > V_{a^c}^*(X_1, \varepsilon_1^*) \mid X]$$

and

$$\Pr[g_a(X_a) + \varepsilon_a \leqslant V_{a^c}(X_1, \varepsilon_1), g_b(X_b) + \varepsilon_b > 0 \mid X, \Psi_2] = \Pr[g_a^*(X_a) + \varepsilon_a^* \leqslant V_{a^c}^*(X_1, \varepsilon_1^*), g_b^*(X_b) + \varepsilon_b^* > 0 \mid X, \Psi_2],$$

$g_a^*$ and $g_b^*$ must be monotonic transformations of $g_a$ and $g_b$. That is it must be the case that for almost any $(x_b^1, x_a^1, x_b^2, x_a^2)$ if

$$g_b(x_b^1) > g_b(x_b^2), \quad g_a(x_a^1) > g_a(x_a^2),$$

then

$$g_b^*(x_b^1) > g_b^*(x_b^2), \quad g_a^*(x_a^1) > g_a^*(x_a^2).$$

I first present the conditions required for identification, pose the theorem, and then describe the general strategy of the proof. The notation $\text{supp}\{Y\}$ denotes the support of random variable $Y$. Since $X_a$ is measurable with respect to $X_1$, the notation $X_a(X_1)$ denotes that value of $X_a$ consistent with $X_1$.

*Condition G1*. For any $x_b \in \text{supp}\{X_b\}$ and $x_1 \in \text{supp}\{X_1\}$,

$$\text{supp}\{\varepsilon_a\} = (S^{\text{l}}_{\varepsilon_a}, S^{\text{u}}_{\varepsilon_a}) \subset \text{supp}\{-g_a(X_a) \mid X_b = x_b)\},$$
$$\text{supp}\{\varepsilon_b\} = (S^{\text{l}}_{\varepsilon_b}, S^{\text{u}}_{\varepsilon_b})$$

($S^{\text{l}}_{\varepsilon_a}, S^{\text{u}}_{\varepsilon_a}, S^{\text{l}}_{\varepsilon_b}$, and $S^{\text{u}}_{\varepsilon_b}$ need not be finite)

*Condition G2*. For any $x_a \in \text{supp}\{X_a\}$, $y \in (-S^{\text{l}}_{\varepsilon_b}, -S^{\text{u}}_{\varepsilon_b})$, and $c \in (0,1)$, there exists a set $\mathscr{X}_1(x_a, y, c)$ with positive measure such that for $x_1 \in \mathscr{X}_1(x_a, y, c)$,

(a) $x_a = X_a(x_1)$,
(b) $\Pr(g_b < y \mid X_1 = x_1) > c$,
(c) The distribution of $g_b$ conditional on $x_1$ is stochastically dominated by the unconditional distribution of $g_b$.

*Condition G3*.

$$\text{E}(|\varepsilon_b| \mid \varepsilon_1) < \infty \quad \text{and} \quad \text{E}(|g_b(X_b)| \mid X_1) < \infty.$$

*Theorem 1. Under Assumptions G1, G2, and G3, $g_a$ and $g_b$ are identified up to monotonic transformations within $(-S^{\text{u}}_{\varepsilon_a}, -S^{\text{l}}_{\varepsilon_a})$ and $(-S^{\text{u}}_{\varepsilon_b}, -S^{\text{l}}_{\varepsilon_b})$ respectively (Proof in Appendix).*

The basic strategy used in this proof is a stochastic extension of 'identification at infinity'. This type of approach is common in static models (see, e.g. Chamberlain (1986), Heckman (1990), Matzkin (1993), or Cameron and Heckman (1998)) and is very difficult to avoid in these types of selection models without parametric restrictions on the distribution of the unobservables. To see how this type of approach works and why it is almost necessary, consider a standard selection model where,

$$d = 1(Z + \varepsilon > 0), \quad y = \beta_0 + u,$$

$\text{E}(u \mid Z) = 0$, $Z$ and $d$ are observable, but $y$ is observed only when $d = 1$. Consider identification of $\beta_0$. If we could condition on a value $z^*$ large enough so that $\Pr(d=1 \mid Z=z^*)=1$ then we could identify $\beta_0$ since $\mathrm{E}(Y \mid Z=z^*)=\beta_0$.[6] We could then trace out the joint distribution of $(\varepsilon, u)$ by varying $Z$.

Two assumptions are important for this strategy. (1) We need an exclusion restriction (a variable $Z$) that enters the selection equation, but not the regression equation, and (2) this variable must have a large support. To see why this second assumption is hard to avoid suppose the support of $Z$ is bounded above by $\bar{Z}$ where $\Pr(d=1 \mid Z=\bar{Z})<1$. In this case for any $\varepsilon < -\bar{Z}$, $d=0$ and $y$ is unobserved. This means that the data is completely uninformative about $\mathrm{E}(u \mid \varepsilon < -\bar{Z})$. Without information about this object, the assumption $\mathrm{E}(u \mid Z)=0$ will not suffice to identify $\beta_0$.[7] To achieve nonparametric identification of $\beta_0$ without placing strong conditions on the conditional distribution of $u$, some type of 'identification at infinity' strategy cannot be avoided.

My model has a similar selection structure. The econometrician can only observe the decision between $b$ and $c$ for individuals who reject $a$. The same intuition for identification that comes from the standard selection model will hold in this case. We typically possess less information in a discrete choice model than in a selection model so it is very difficult to avoid the 'identification at infinity' strategy here as well without strong restrictions on the error terms.

I identify $g_b$ in almost exactly the same manner as $\beta_0$ in the above example. With an exclusion restriction we can condition on $g_a$ arbitrarily low so that the probability of selecting node $a$ is close to zero. This leaves us with a simple binary choice model in which the agents choose between $b$ and $c$. From previous work we know in this case that we can identify $g_b$ up to a monotonic transformation. The type of exclusion restriction used here is a variable that enters $g_a$, but does not influence $g_b$ directly. To see this suppose that $X_a$ is unidimensional, does not influence $g_b$, and that,

$$\lim_{x_a \to -\infty} g_a = -\infty,$$

then,

$$\begin{aligned}\lim_{x_a \to -\infty} \Pr(b \mid X) &= \lim_{x_a \to -\infty} \Pr[g_a(X_a)+\varepsilon_a \\ &\leqslant V_{a^c}(X_1, \varepsilon_1),\ g_b(X_b)+\varepsilon_b > 0 \mid X] \\ &= \Pr[g_b(X_b)+\varepsilon_b > 0 \mid X].\end{aligned}$$

[6] And similarly if we could condition on $Z$ so that $\Pr(d=1 \mid Z)$ is 'close' to one then we could obtain an estimate 'close' to $\beta_0$.

[7] Heckman and Vytlacil (1999), Aakvik et al. (1999), and Ichimura and Taber (1999) use a different strategy. They consider the case where one has exclusion restrictions for this problem, but not full support of $Z$. In this case one cannot get point estimates of $\beta_0$ but can get bounds on these values. A similar strategy could be used for the model presented here as well.

Using standard identification strategies for the binary choice model (see, e.g. Manski (1988) or Matzkin (1992)), I can identify $g_b$.[8] If we have a variable that influences $g_a$, but not $g_b$ directly then we can fix $X_b$ and still vary $g_a$. This type of exclusion restriction satisfies G1. Note that time varying $X$'s are typically sufficient for an exclusion restriction here. A first period outcome will influence $g_a$, but not influence $g_b$ conditional on the second period outcome.

Identification of $g_a$ is somewhat trickier. Since the sequencing of the choices is different, at first glance the problem does not seem to take the form of the selection model. However, it is similar. Since $V_c$ is normalized to zero, $g_a$ represents the difference in utility between $a$ and $c$ that is made given information at time 1. If we could condition on a group of people for whom $b$ is not an option, then we could identify $g_a$ using the same argument as above. Since in general $g_b$ will depend on values of $X_b$ that are not realized until time two we cannot condition on $g_b$ at time one. Instead I develop a stochastic notion of identification at infinity. Rather than conditioning on a set of $X_b$ such that $g_b$ is small, I condition on a set of $X_1$ such that the conditional distribution of $g_b$ is 'small.'

This requires a somewhat different type of exclusion restriction, a variable known at time one that does not enter $g_a$ directly, but does have predictive power for the distribution of $g_b$ above and beyond $X_a$. To see how this works, suppose we have a variable $X_1$ that satisfies these conditions and that as $x_1$ gets small the conditional distribution of $g_b$ becomes small. In this case

$$\lim_{x_1 \to -\infty} \mathrm{E}[\max(g_b, 0) \mid x_1, \varepsilon_1] = 0,$$

so that

$$\begin{aligned}\lim_{x_1 \to -\infty} \Pr(a \mid X) &= \lim_{x_1 \to -\infty} \Pr[g_a(x_a) + \varepsilon_a > \mathrm{E}[\max(g_b, 0) \mid x_1, \varepsilon_1]] \\ &= \Pr[g_a(x_b) + \varepsilon_a > 0].\end{aligned}$$

From this piece we can identify $g_a$ up to a monotonic transformation. This type of variable will satisfy G2. Note that simple time varying $X$'s will not typically be sufficient in this case. We need a variable that is known at time one and does not enter $g_a$ directly. A second period realization of an observable will not enter $g_a$ directly, but it typically will not be known at time one.

[8] I assumed one piece of information that is not available. I assumed that the econometrician knew that $g_a$ went to negative infinity with $x_a$ even though I have not shown that $g_a$ is identified. This is not a serious issue since holding $g_b$ constant, the set of $x_a$ for which $g_a \to -\infty$ is the same as the set of $x_a$ for which the probability of choosing $a$ goes to zero which is observable.

Given a set of exclusion restrictions with large enough support, the model is essentially transformed from a dynamic model to a static binary choice model. Thus, the identification strategy here can be easily extended to other cases addressed in that literature. For example, if one wanted to allow for heteroskedasticity in the error terms as in Manski (1975), a combination of Assumptions G1 and G2 as well as Manski's assumptions would be sufficient for identification. Extending the model to more periods and more choices is also straight forward. With multiple choices one needs multiple exclusion restriction that would be jointly sent to infinity. Once again it would be almost impossible to nonparametrically identify the model without this type of assumptions. Extending the model to allow for endogenous continuous variables in a manner similar to Heckman and Honoré (1990) was done in Taber (1996). It also uses the intuition presented here.

The assumptions above about access to exclusion restrictions can be relaxed if one is willing to make parametric assumptions about $g_a$ and $g_b$. In particular if $g_a = X'\beta_a$ and $g_b = X'\beta_b$, then exclusion restrictions are no longer necessary (see Taber, 1996). To see the intuition for this, as long as $\beta_a$ is not proportional to $\beta_b$ we can send $g_a \to \infty$ , and still have enough variation in $X$ to identify $\beta_b$.

## 4. Identification of the distribution of the unobservables

The theorems above show that $g_a$ and $g_b$ can be estimated even if we can say nothing about the distribution of the error terms. However, their nonparametric identification is of interest as well. Typically these types of structural models are estimated with the goal of simulating policy counterfactuals. Except in very special cases, without knowledge of the full model, these counterfactuals cannot be constructed. For example, in the schooling case, suppose that policy makers consider subsidizing college education. Evaluating the consequences of the policy on schooling outcomes from the model cannot be done using $g_a$ and $g_b$ alone. A second reason for exploring identification of the distribution of unobservables is that nonparametric identification of the unobservables is required for the use of many semiparametric estimators. For example, showing consistency of the nonparametric maximum likelihood estimator that I use below requires identification of the distribution of the unobservables. Finally, the joint distribution of the unobservables may be of interest in its own right. In the schooling model a researcher may be interested in understanding the manner in which students learn about their own ability.

The most general version of the full model above cannot be identified without further assumptions. I will consider the following possible restrictions on the unobservables that may provide identification.

*Assumption E1*. For all $y \in \Re$, $\Pr[\varepsilon_b \leqslant y \mid \varepsilon_1, \Psi_2] = \Pr[\varepsilon_b \leqslant y \mid \varepsilon_1]$.

*Assumption E2*. For all $y \in \Re$, $\Pr[\varepsilon_b \leqslant y \mid \varepsilon_1] = \Pr[\varepsilon_b \leqslant y \mid \varepsilon_a]$.

*Assumption E3*. $\varepsilon_b = v_b + \eta_b$ where $v_b = \mathrm{E}(\varepsilon_b \mid \varepsilon_1)$, $\eta_b$ is independent of $\varepsilon_1$, and the characteristic functions of $\varepsilon_b$ and $\eta_b$ do not vanish.

Assumption E1 eliminates the possibility of a difference between the ex-ante and ex-post conditional distribution of $\varepsilon_b$. It is helpful for identification because it places strong restrictions on the relationship between the conditional distribution of $\varepsilon_b$ and the conditional distribution agents possess about $\varepsilon_b$ during the first time period. Without this assumption, or at least a strong restriction on the way these effects operate, identification of the full model from only one realization of $\Psi_2$ is not feasible.

Assumptions E2 and E3 are alternative conditions on the first period information people possess about their unobservables. Neither is stronger than the other. Assumption E2 essentially allows for general types of serial correlation. It imposes that agents have no information about $\varepsilon_b$ beyond $\varepsilon_a$, but does not restrict the relationship between $\varepsilon_a$ and $\varepsilon_b$. Assumption E3 allows a very general conditioning set, but restricts the knowledge of $\varepsilon_b$ to be simply its expected value. In the schooling model one might expect that individuals have more information about their returns to college than is conveyed through their returns to high school, so Assumption E3 is probably more appropriate. However, in cases in which the decision between $a$ and $a^c$ is similar to the choice between $b$ and $c$, Assumption E2 may be preferred.

The first result of this section is that even under the seemingly strong conditions above, identification of the error structure cannot be achieved without stochastic innovations in the observables between the two periods. That is, when agents know the value of $X_b$ with perfect certainty in the first period, identification cannot be achieved even under strong parametric assumptions. The basic problem is that the choice in the first period is influenced by $V_{a^c}(X_1, \varepsilon_1)$. If $X_b$ were known with perfect certainty in the first period then $X_1 = X_b$ and we could not vary $g_b(X_b)$ separately from $V_{a^c}(X_b, \varepsilon_1)$. Under Assumption E3, that would leave us with essentially two degrees of freedom $(g_a, g_b)$ to identify a three dimensional distribution $(\varepsilon_a, v_b, \eta_b)$. Under Assumption E2 the intuition is more subtle. Since $g_b(X_b)$ enters both the first and second period decisions, it is not possible to differentiate between the two roles which is necessary for identification in some cases.

I first use counterexamples to demonstrate nonidentification of the distribution of the error terms in this case. I then show that with stochastic innovations in $X_b$, I can vary $V_{a^c}(X_1, \varepsilon_1)$ separately from $g_b(X_b)$ which delivers identification of the distribution of the error terms under condition E1 and either E2 or E3. While these counterexamples are very special, only very restrictive general conditions will rule them out. Unless we use these other very strong assumptions, stochastic innovations in the observables between periods are necessary for identification.

In what follows I assume that $g_a$ and $g_b$ are identified. In the first section I showed that they could be identified up to monotonic transformations. Therefore, after choosing a class of functions which are normalized up to a monotonic transformation, they are identified. There are a number of different normalizations have been used in the binary choice model that can be used here as well (see, e.g. Manski (1988), Cosslett (1983) or Matzkin (1990,1992)). I will not discuss specific ones but refer the reader to previous work. The only somewhat unique aspect of this problem is that we can only normalize one of these functions, and given this normalization the other should be identified. For example in the linear case if we normalize the scale of $g_a$ we can identify the scale of $g_b$ under the conditions presented in the previous section. In some cases when $g_a$ and $g_b$ are completely nonparametric this identification requires an additional exclusion restrictions (a variable that influences $g_c$, but not $g_a$ or $g_b$ directly). These issues are much easier to deal with under specific forms of $g_a$ and $g_b$ rather than in the general case, so for the sake of space, I just assume these conditions hold rather than get into these details.

*Assumption G4.* $g_a(X_a)$ and $g_b(X_b)$ are identified.

I first consider the case in which $X_b$ is known to the agent with perfect certainty during the first period, so $\mathrm{E}(g_b \mid X_1) = g_b$. Notice that when $g_b \to -\infty$, $\Pr(a) \to \Pr(g_a + \varepsilon_a > 0)$, so we can identify the distribution of $\varepsilon_a$. Similarly if we set $g_a \to -\infty$, $\Pr(b) \to \Pr(g_b + \varepsilon_b > 0)$, so we can identify the distribution of $\varepsilon_b$. The problem is that we cannot identify the joint distribution.

I first show through a counterexample that Assumptions E1 and E2 are not sufficient for identification in the case where $X_b$ is known during the first period. The basic intuition is that we do not have enough variation in the observables to separate the direct effect of $\varepsilon_a$ from its role in predicting $\varepsilon_b$.

*Counterexample 1.* Assume that $\varepsilon_a$ is binomial and that the distribution $\varepsilon_b$ conditional on $\varepsilon_a$ is also binomial for each value of $\varepsilon_a$. I let the $(\varepsilon_a, \varepsilon_b)$ have the following distribution,

$$\varepsilon_a = \begin{cases} \theta_1 & \text{with probability } \rho, \\ \theta_2 & \text{with probability } 1-\rho, \end{cases}$$

$$(\varepsilon_b \mid \varepsilon_a = \theta_1) = \begin{cases} -\phi_1 & \text{with probability } \mu, \\ -\phi_a & \text{with probability } 1-\mu, \end{cases}$$

$$(\varepsilon_b \mid \varepsilon_a = \theta_2) = \begin{cases} -\phi_2 & \text{with probability } \mu, \\ -\phi_b & \text{with probability } 1-\mu, \end{cases}$$

where $\phi_a > \phi_1$ and $\phi_b > \phi_2$. If $\rho = 0.5$ and $\theta_1 - \phi_1\mu = \theta_2 - \phi_2\mu$, then the model with $\phi_a = \phi_3$ and $\phi_b = \phi_4$ cannot be distinguished from an alternative model with $\phi_a = \phi_4$ and $\phi_b = \phi_3$.

Now consider Assumption E3. I will go to the two extremes and provide a counterexample in which I cannot distinguish a model in which the agent has full knowledge of $\varepsilon_b$ during the first period (i.e. $\varepsilon_b = v_b$) from a model in which the agent has no knowledge of $\varepsilon_b$ during the first period (i.e. $\varepsilon_b = \eta_b$). I take $\varepsilon_a$ and $\varepsilon_b$ to be distributed logistically and show that the nested logit model cannot be distinguished from a model in which agents have no information about $\varepsilon_b$ at time one. McFadden has shown that the nested logit can be derived from a multinomial choice model. These models are special cases in which the agents have full information in the first period.

*Counterexample 2*. I present the models in the context of my current notation without the normalization of $g_c = 0$. I let $\hat{g}_k$ be the original reward functions so by definition $g_a = \hat{g}_a - \hat{g}_c$ and $g_b = \hat{g}_b - \hat{g}_c$. The following two models produce the same choice probabilities.

Model 1 (Nested Logit Model(McFadden 1977,1981))

$$V_a = \hat{g}_a + \hat{\varepsilon}_a,$$

$$V_b = \hat{g}_b + \hat{\varepsilon}_b,$$

$$V_c = \hat{g}_c + \hat{\varepsilon}_c,$$

$$\mathscr{I}_1 = \sigma(\hat{g}_a, \hat{g}_b, \hat{g}_c, \hat{\varepsilon}_a, \hat{\varepsilon}_b, \hat{\varepsilon}_c),$$

$$F(\hat{\varepsilon}_a, \hat{\varepsilon}_b, \hat{\varepsilon}_c) = \exp(-\exp(-\hat{\varepsilon}_a))\exp\left(-\left[\exp\left(\frac{-\hat{\varepsilon}_b}{\rho}\right) + \exp\left(\frac{-\hat{\varepsilon}_c}{\rho}\right)\right]^{\rho}\right)$$

Model 2:

$$V_a = \hat{g}_a + \hat{\varepsilon}_a,$$

$$V_b = \hat{g}_b + \omega + \rho\hat{\varepsilon}_b,$$

$$V_c = \hat{g}_c + \omega + \rho\hat{\varepsilon}_c,$$

$$\mathscr{I}_1 = \sigma(\hat{g}_a, \hat{g}_b, \hat{g}_c, \hat{\varepsilon}_a, \omega),$$

$$F(\hat{\varepsilon}_a, \hat{\varepsilon}_b, \hat{\varepsilon}_c, \omega) = \exp(-\exp(-\hat{\varepsilon}_a))\exp(-\exp(-\hat{\varepsilon}_b))$$
$$\exp(-\exp(-\hat{\varepsilon}_c))\exp(-\exp(-\omega)).$$

In other words the error terms are all independent with Type 1 extreme value distribution.

Now suppose that $X_b$ is not known with perfect certainty during the first period. In this case it is possible to provide sufficient conditions under which the distribution of the unobservables is identified. I assume E1 and show that either E2 or E3 are sufficient for identification. I use the following additional assumption,

*Condition G5*. For almost all $x_1 \in \mathrm{supp}(X_1)$, $(S^{\mathrm{l}}_{\varepsilon_b}, S^{\mathrm{u}}_{\varepsilon_b}) \in \mathrm{supp}(-g_b(X_b) \mid X_1 = x_1)$.

I first show in the following lemma that this additional assumption provides identification of the joint distribution of $(\varepsilon_a, \varepsilon_b)$. I then use this lemma to prove I can identify the full model when I combine Assumption E1 with either E2 or E3.

*Lemma 1. Under Assumptions G1–G5 the joint distribution of* $(\varepsilon_a, \varepsilon_b)$ *is identified.* (*Proof in Appendix.*)

To see the intuition for the proof of the lemma recall that,

$$\Pr(b \mid X) = \Pr(g_b + \varepsilon_b > 0, g_a + \varepsilon_a \leqslant V_{a^c}(X_1, \varepsilon_1) \mid X).$$

So by sending $V_{a^c}(X_1, \varepsilon_1) \to 0$ as in the proof of the first theorem, I can identify

$$\Pr(g_b + \varepsilon_b > 0, g_a + \varepsilon_a \leqslant 0 \mid X),$$

from which it is easy to identify the joint distribution of $(\varepsilon_a, \varepsilon_b)$ by varying $g_a$ and $g_b$.

Given this lemma it is obvious that Assumptions E1 and E2 are sufficient for identification.

*Theorem 2. Under Assumptions E1, E2, and G1–G5 the full model is identified* (*Proof in Appendix.*)

I now consider Assumption E3. This is useful because as $\mathrm{E}(g_b + v_b \mid X_1, \varepsilon_1)$ gets large, $\mathrm{E}[\max\{g_b + v_b + \eta_b, 0\} \mid X_1, v_b]$ approaches $\mathrm{E}(g_b \mid X_1) + v_b$. I use this fact to show that I can identify the joint distribution of $(\varepsilon_b - v_b, v_b + \eta_b)$, and from this I can identify the distribution of $\eta_b$ and the joint distribution of $(\varepsilon_b, v_b)$.

*Theorem 3. Under Assumptions E1, E3, and G1–G5 the full model is identified.* (*Proof in Appendix.*)

## 5. Estimation of a schooling model

In this section I estimate an empirical schooling model using the framework developed above. There is a very large literature in labor economics, public economics, and sociology on schooling decisions. Perhaps the largest concern in this literature has been about heterogeneity and selection bias. In terms of observable attributes, students who attend college are very different than those who do not. It is thus reasonable to expect that they are different in terms of unobservable attributes as well. Cameron and Heckman (1998) provide a recent example of a schooling model that focuses on heterogeneity and Card (1998) provides a recent survey of work done on the returns to schooling which deal with the selection problem in a variety of ways. Schooling is also clearly a dynamic decision in which people do not have full certainty about their options when they make the decisions. Many papers in this literature have addressed this problem of uncertainty in schooling returns. Examples include Weisbrod (1962), Comay et al. (1973), Altonji (1993), Belzil and Hansen (1997), Keane and Wolpin (1997), Buchinsky and Leslie (1996), and Taber (1998). In this section I apply the discussion of identification above to a dynamic schooling model. Given the exclusion restrictions suggested by the assumptions above, I estimate a version of the model.

To be consistent with the simple framework above, I consider two schooling decisions, the first is whether to graduate from high school, and the second is whether to attend college. At the time the high school graduation decision is made, students do not know with perfect certainty whether they will attend college. In terms of the notation above node $a$ represents dropping out of high school, node $a^c$ graduating from high school, node $c$ entering college, and node $b$ entering the labor force immediately following high school graduation. The model I estimate is a modified version of the specification developed in Cameron and Taber (1998) and details about the data can be found there. In the previous section, I presented two possible manners of representing the information structure, E2 and E3. For the schooling model, Assumption E3 seems more appropriate. As discussed above, typically we think that high school students will have some private information about their own returns to schooling that is known during high school. The serial correlation Assumption E2 does not capture this very well since the determinants of college matriculation may depend on different attributes than that for high school graduation. For example the types of skills that are relatively more important for college sector jobs than high school graduate sector jobs, seem very different than the type of skills that are relatively more important for high school sector jobs versus high school dropout jobs. Under this assumption as well as linearity, the value functions take the form,

$$V_a = X_a'\beta_a + \varepsilon_a,$$

$$V_b = X_b'\beta_b + v_b + \eta_b.$$

Assumption E3 is also restrictive. It assumes that while $v_b$ is known during high school, there is no variation in the conditional variance of the agent's forecast of $V_b$.

Previous empirical work on selection models has found that empirical estimates are much more reliable when exclusion restrictions are used for identification even in parametric cases in which exclusion restrictions are not necessary. There is a huge literature on the returns to schooling that considers different exclusion restrictions.[9] The results above suggest that two types are likely to be useful. (1) Assumptions G1 and G5 can be satisfied with time varying $X$'s. (2) To satisfy Assumption G2, I need a variable known at time one, that influences the decision about whether to attend college, but does not affect the returns to dropping out of high school directly.

Local labor market variation provides a potential source of time varying observables. Temporarily low wage rates will lower the opportunity cost of schooling and lead more individuals to attend college. The local wage during high school satisfies the type of exclusion restriction needed for G1: it is a variable that influences the decision to drop out of college, but conditional on the local wage during college, should have no effect on the college decision. Assumption G5 requires a variable that is not known at time one, but influences the time two decision. The college local wage variable satisfies this condition. It influences the college decision, but is not known with perfect certainty during the first period. Measures of the cost of college will satisfy Assumption G2 exclusion restrictions, they are often known during high school, but should have no direct effect on high school graduation. I use a dummy variable for whether there is a college in the student's county. This should certainly influence the probability of attending, will be known during high school, but should have no direct effect the decision to drop out. While these variables do seem to satisfy the criterion for exclusion restrictions, they do not have large support so they are not ideal.

I estimate the model using a flexible form for the distribution of the error terms. In particular, I assume that I can write $\varepsilon_a = \varepsilon_a^1 + \varepsilon_a^2$ where $\varepsilon_a^2$ is standard normal and independent of $v_b$. I estimate the distribution of $(\varepsilon_a^1, v_b)$ by assuming these variables take on only finitely many values. By letting the number of values get large I can approximate any smooth distribution function arbitrarily well. Heckman and Singer (1984) show consistency of a similar procedure and along with Cameron and Taber (1998) have monte carlo results that demonstrate that this approximation works very well in practice. Similarly, I assume that $\eta_b = \eta_b^1 + \eta_b^2$ where $\eta_b^2$ is independent of $\eta_b^1$ and is normal mean zero, and that $\eta_b^1$ takes only finitely many values. Specifically, $(\varepsilon_a^1, v_b)$ takes $K_1$ values which I denote by $(\varepsilon_{aj_1}^1, v_{bj_1})$ each with probability $\mu_{1j_1}$ for $j_1 = 1, \ldots, K$. Similarly

[9] Again, Card (1998) provides a good survey.

$\eta_b^1$ takes $K_2$ values denoted by $\eta_{bj_2}^1$ each with probability $\mu_{2j_2}$ for $j_2 = 1, \dots, K_2$. Under this notation,

$$V_{a^c}(X_1, v_b) = \mathrm{E}(\max\{X_b'\beta_b + v_b + \eta_b^1 + \eta_b^2, 0\} \mid X_1, v_b)$$

$$= \int \sum_{j_2=1}^{K_2} \sigma_b \varphi\left(\frac{X_b'\beta_b + v_b + \eta_{bj_2}^1}{\sigma_b}\right)\mu_{2j_2}\, \mathrm{d}G(X_b \mid X_1),$$

where

$$\varphi(Y) = \Phi(Y)(Y) + \phi(Y),$$

$\sigma_b$ is the standard deviation of $\eta_b^2$, and $\Phi$ and $\phi$ denote the CDF and PDF of a standard normal random variable. We can then form the pieces of the likelihood function given,

$$\Pr(d_a = 1 \mid X_1) = \Pr(X_a'\beta_a + \varepsilon_a^1 + \varepsilon_a^2 > V_{a^c}(X_1, v_b) \mid X_1)$$

$$= \sum_{j_1=1}^{K_1} \Phi(X_a'\beta_a + \varepsilon_{aj_1}^1 - V_{a^c}(X_1, v_{bj_1}))\mu_{1j_1},$$

and

$$\Pr(d_b = 1 \mid X_1, X_b)$$

$$= \sum_{j_1=1}^{K_1} \Bigg[(1 - \Phi(X_a'\beta_a + \varepsilon_{aj_1}^1 - V_{a^c}(X_1, v_{bj_1})))$$

$$\left(\sum_{j_2=1}^{K_2} \Phi\left(\frac{X_b'\beta_b + v_{bj_1} + \eta_{bj_2}^1}{\sigma_b}\right)\mu_{2j_2}\right)\Bigg]\mu_{1j_1}.$$

For any given $K_1$ and $K_2$ I use maximum likelihood, estimating the parameters

$$[\beta_a, \beta_b, \sigma_b, (\varepsilon_{a1}^1, v_{b1}), \dots, (\varepsilon_{aK_1}^1, v_{bK_1}), \eta_{b1}^1, \dots, \eta_{bK_2}^1].$$

I present a model estimated without heterogeneity in Table 1. Included in the specification are family background variables, test scores from four sections of the Armed Service Vocational Aptitude Battery test administered to individuals in the NLSY sample, demographic variables, regional variables, cohort dummies, and the exclusion restrictions. Most variables enter the model as in previous work. Of particular concern are the exclusion restrictions. As expected having a college in one's county does make individuals more likely to attend college. I want to use the local cyclical patterns of local wages for identification rather than cross-sectional differences, as they may be attributed instead to differences in wealth levels across counties. With this in mind I control for the

Table 1
Estimated parameters from schooling model with no heterogeneity

| Variables | Drop out of high school | Attend college |
|---|---|---|
| Constant | 1.444 (0.230) | − 1.412 (0.268) |
| Live in South | 0.109 (0.106) | 0.387 (0.095) |
| Live in West | 0.035 (0.125) | 0.399 (0.109) |
| Live in Northeast | 0.229 (0.126) | − 0.181 (0.118) |
| Math Score | − 0.082 (0.016) | 0.074 (0.008) |
| Science Score | − 0.023 (0.012) | 0.011 (0.012) |
| Word Score | − 0.023 (0.010) | 0.046 (0.008) |
| Automotive Knowledge Score | − 0.008 (0.012) | − 0.048 (0.009) |
| Highest Grade Compl. Father | − 0.020 (0.013) | 0.040 (0.013) |
| Highest Grade Compl. Mother | − 0.002 (0.017) | 0.039 (0.017) |
| Number of Siblings | 0.013 (0.014) | − 0.021 (0.015) |
| Black | − 0.415 (0.105) | − 0.324 (0.101) |
| Hispanic | − 0.066 (0.122) | 0.398 (0.117) |
| College in County | | 0.363 (0.102) |
| Average Wage in County | 0.024 (0.071) | 0.230 (0.072) |
| Wage in County at Time | 0.039 (0.081) | − 0.235 (0.080) |
| Cohort Dummies | Yes | Yes |
| Standard error of $\varepsilon_b$ | 0.330 (0.328) | |

Note: Standard errors in parentheses.

long-run mean wage in the county over an approximately thirty-year period. The level of average wages at age 16 enters the decision about whether to drop out of high school and the level of average wages at age 18 enters the decision about whether to attend college.[10] These variables have the expected signs in the college decision.[11] Students from counties with higher average income are more likely to attend college, and college attendance is counter-cyclical. Unfortunately, the local labor market variables are much weaker in the high school drop out decision. It is also notable that the standard error of $\varepsilon_b$ is not significant in this model. Looking at the probabilities above we see that this parameter is essentially the coefficient on

$$\int \left[ \Phi\left(\frac{X_b'\beta_b}{\sigma_b}\right)\left(\frac{X_b'\beta_b}{\sigma_b}\right) + \phi\left(\frac{X_b'\beta}{\sigma_b}\right) \right] \mathrm{d}G(X_b \mid X_1)$$

[10] In order to avoid endogeneity associated with moving, both the local wages and the college in county are measured based on where the respondent lived at age 17.

[11] I approximated $G(X_b \mid X_1)$ by assuming that the log deviation of the local labor market variable from its long run mean follows an AR(1) with a gaussian error term. This approximation seems to fit the data well.

in a probit for whether the individual drops out of high school. The exclusion restriction that identifies this parameter is the 'college in county' dummy variable. A reduced form probit on high school drop out gives a very similar result, the coefficient on this variable is positive but not significant. There are a number of different interpretations of this result that the option value of college does not seem to affect high school completion. The first is that we simply need more data to get a better estimate of the effect. A second is that this is evidence that high school students are not forward looking. A third is that the option of college has no value to high school dropouts. That is, it is possible that individuals at the margin of whether to drop out of high school, would not attend college if they did complete high school. For them, the cost of college is irrelevant so the decision about whether to drop out of high school will not be influenced by college costs. Distinguishing between these three possibilities is beyond the scope of this paper.

I next present results from the specification in which I allow heterogeneity to enter the model flexibly. The basic strategy is to add points of support to the distribution of the heterogeneity until the likelihood fails to increase by some prespecified amount. In particular, I use the Akaike Information Criterion to choose the number of points of support. The final model gave me a value of $K_1 = 5$ and $K_2 = 0$.[12] The results of this model are presented in Table 2. There are a few strange aspects to the results. The most striking is the size of the coefficients and the support points for the heterogeneity distribution in the college attendance decision. With these estimates, the variance of $v_b$ is very large relative to the variance of $\eta_b$. $\eta_b$ is essentially irrelevant as a predictor of college attendance. If the variance of $\eta_b$ were zero, the model would not be differentiable and the standard method of approximating standard errors would not work. While this is not precisely true here, with these estimates it is approximately true so the estimates of the standard errors are not likely to be reliable. Most coefficients in the schooling decision have the expected signs, but nothing is close to being statistically significant at conventional levels. We also see that the standard error of $\varepsilon_b$ once again is insignificant and in this case has the wrong sign. Given the large and unreliable standard errors it is difficult to make strong claims about the interpretation of these results. To be able to estimate the dynamics of high school completion, more work needs to be done with hopefully more powerful exclusion restrictions, though finding such covariates may be very difficult.

## 6. Summary and conclusions

This paper develops a simple discrete choice dynamic programming model with a quite general form for unobservables and agent's information sets. The

[12] At essentially every level of $K_1$ I found no evidence that $K_2$ should be higher than zero.

Table 2
Estimated parameters from schooling model with five points of support

| Variables | Drop out of high school | Attend college |
|---|---|---|
| Constant | 1.236 (0.458) | 4313.610 (10452.171) |
| Live in South | 0.043 (0.155) | 524.178 (3466.992) |
| Live in West | 0.012 (0.191) | 715.894 (4722.897) |
| Live in Northeast | 0.278 (0.190) | 78.88 (4339.117) |
| Math Score | − 0.109 (0.019) | 105.399 (272.429) |
| Science Score | − 0.029 (0.019) | 31.834 (382.258) |
| Word Score | − 0.035 (0.013) | 85.417 (267.042) |
| Automotive Knowledge Score | 0.005 (0.016) | − 99.152.942 (312.035) |
| Highest Grade Compl. Father | − 0.030 (0.017) | 59.370 (309.565) |
| Highest Grade Compl. Mother | − 0.019 (0.025) | 111.596 (432.254) |
| Number of Siblings | 0.018 (0.021) | 12.135 (416.229) |
| Black | − 0.507 (0.149) | 321.498 (2636.381) |
| Hispanic | − 0.147 (0.170) | 547.318 (2912.296) |
| College in County | | 79.928 (3219.761) |
| Average Wage in County | 0.018 (0.019) | 218.244 (1875.892) |
| Wage in County at Time | 0.046 (0.099) | − 216.574 (2091.232) |
| Cohort Dummies | Yes | Yes |
| Distribution of heterogeneity | | |
| Probability | Drop out of high school | Attend college |
| 0.317 (174.143) | 0.000 | 0.00 |
| 0.392 (142.507) | 0.000 | 1728.09 |
| 0.166 (25.754) | − 0.315 | 3857.79 |
| 0.069 (4.681) | 392.307 | 3234.37 |
| 0.056 (1.120) | 392.700 | − 2599.65 |
| Standard error of $\varepsilon_b$ | | − 0.00008 (0.000020) |

Note: Standard errors in parentheses.

goal is to uncover what type of data can solve the selection problem induced by this structure. As in static models, I show that with strong support conditions and exclusion restrictions the model is identified. While these support conditions are strong, it is very difficult to avoid them. Essentially two types of exclusion restriction are required. The first is a variable that influences the first period decision, but does not enter the second period decision directly. The second type requires a variable that does not affect the utility of the first option directly, but is known during the first period and has predictive power on the choice during the second. I also provide two specifications under which the full error structure can be identified. This requires the additional assumption of stochastic innovations in the $X$'s: a variable known at time one that helps predict the second period decision, but conditional on second period observables, has no influence on the decision. While the model presented here is special, generalizing these results to more complicated finite time models is straight forward.

I estimate a schooling version of the model in which students first decide whether to graduate from high school and then decide whether to attend college. This procedure has only limited success. The model does not show signs of forward looking behavior and reliable standard errors could not be obtained. Part of the problem may be that the exclusion restrictions are weaker than one may hope, and they do not have large support. One possible direction for future research on dynamic schooling models is to obtain more powerful exclusion restrictions which may solve the problems, although this may prove difficult. More generally this paper has suggested that certain types of exclusion restrictions with strong support conditions should help solve the dynamic selection problem. This should be a useful input for empiricists who face this problem.

**Acknowledgements**

I would like to thank Steve Cameron, Tim Conley, Bo Honoré, Joe Hotz, Hide Ichimura, Rosa Matzkin, Chuck Manski, Ariel Pakes, several referees, and especially James Heckman for helpful comments. An earlier version of this work was part of my dissertation at the University of Chicago. I gratefully acknowledge the financial support of the Searle Foundation and the Alfred P. Sloan Foundation. All remaining errors are my own.

## Appendix

*Proof of Theorem 1.* Since every probability I consider in this section conditions on $X$ and $\Psi_2$, for the sake of exposition I leave this conditioning implicit.

Suppose that there exists $(g_a, g_b) \neq (g_a^*, g_b^*)$, $\varepsilon(\omega)$ and $\varepsilon^*(\omega)$ such that for almost all $(x_1, x_a, x_b)$,

$$\Pr[g_a(x_a) + \varepsilon_a > V_{a^c}(x_1, \varepsilon_1)] = \Pr[g_a^*(x_a) + \varepsilon_a^* > V_{a^c}^*(x_1, \varepsilon_1^*)] \tag{A.1}$$

and

$$\Pr[g_a(x_a) + \varepsilon_a \leqslant V_{a^c}(x_1, \varepsilon_1), g_b(x_b) + \varepsilon_b > 0]$$

$$= \Pr[g_a^*(x_a) + \varepsilon_a^* \leqslant V_{a^c}^*(x_1, \varepsilon_1^*), g_b^*(x_b) + \varepsilon_b^* > 0]. \tag{A.2}$$

I will first show that $g_b^*$ must be a monotonic transformation of $g_b$ on the limited support.

Suppose not, suppose there exist $\mathscr{X}_b^1$ and $\mathscr{X}_b^b$ with positive measure such that for all $x_b^1 \in \mathscr{X}_b^1$ and all $x_b^2 \in \mathscr{X}_b^2$,

$$-S_{\varepsilon_b}^{\mathrm{u}} > g_b(x_b^1) > g_b(x_b^2) > -S_{\varepsilon_b}^{\mathrm{l}}$$

$$g_b^*(x_b^1) < g_b^*(x_b^2),$$

then for $\delta$ small enough, from the conditions on the support $\varepsilon_b$ either for $x_b^1 \in \mathscr{X}_b^1$,

$$\Pr[g_b(x_b^1) + \varepsilon_b > 0] - \Pr[g_b^*(x_b^1) + \varepsilon_b^* > 0] > \delta,$$

or for $x_b^2 \in \mathscr{X}_b^2$,

$$\Pr[g_b^*(x_b^2) + \varepsilon_b^* > 0] - \Pr[g_b(x_b^2) + \varepsilon_b > 0] > \delta.$$

Without loss of generality suppose it is the first. From Condition G1 we can find a $X_a(x_b^1)$ such that,
$\Pr[g_a(X_a(x_b^1)) + \varepsilon_a > 0] < \delta$. Then for all $x_b^1 \in \mathscr{X}_b^1$,

$$\begin{aligned}
\delta &> \Pr[g_a(X_a(x_b^1)) + \varepsilon_a > 0] \\
&\geqslant \Pr[g_a(X_a(x_b^1)) + \varepsilon_a > V_{a^c}(x_1, \varepsilon_1), g_b(x_b^1) + \varepsilon_b > 0] \qquad \text{(A.3)} \\
&= \Pr[g_b(x_b^1) + \varepsilon_b > 0] - \Pr[g_b(x_b) + \varepsilon_b > 0, g_a(X_a(x_b^1)) \\
&\quad + \varepsilon_a \leqslant V_{a^c}(x_1, \varepsilon_1)] \\
&\geqslant \Pr[g_b(x_b^1) + \varepsilon_b > 0] - \Pr[g_b(x_b) + \varepsilon_b > 0, g_a(X_a(x_b^1)) \\
&\quad + \varepsilon_a \leqslant V_{a^c}(x_1, \varepsilon_1)] \\
&\quad - [\Pr[g_b^*(x_b^1) + \varepsilon_b^* > 0] - \Pr[g_b^*(X_b) + \varepsilon_b^* > 0, g_a^*(X_a(x_b^1)) \\
&\quad + \varepsilon_a^* \leqslant V_{a^c}^*(x_1, \varepsilon_1^*)]] \\
&= \Pr[g_b(x_b^1) + \varepsilon_b > 0] - \Pr[g_b^*(x_b^1) + \varepsilon_b^* > 0]. \qquad \text{(A.4)}
\end{aligned}$$

which is a contradiction so $g_b$ must be identified to a monotonic transformation on the limited support.

Now in a similar manner suppose that $g_a$ is not identified up to a monotonic transformation. From the same argument as above, there must exist a set $\mathscr{X}_a^1$ of positive measure such that for all $X_a^1 \in \mathscr{X}_a^1$, $-S_{\varepsilon_b}^{\mathrm{u}} > g_b(x_a^1) > -S_{\varepsilon b}^{\mathrm{l}}$ and

$$\Pr[g_a(x_a^1) + \varepsilon_a > 0] - \Pr[g_a^*(x_a^1) + \varepsilon_a^* > 0] > \delta.$$

For any $x_1 \in \operatorname{supp}\{X_1\}$ for which, $X_a(x_1) \in \mathscr{X}_a^1$,

$$\begin{aligned}
\delta &< \Pr[g_a(X_a(x_1)) + \varepsilon_a > 0] - \Pr[g_a^*(X_a(x_1)) + \varepsilon_a^* > 0] \\
&= \Pr[g_a(X_a(x_1)) + \varepsilon_a > 0] - \Pr[g_a(X_a(x_1)) + \varepsilon_a > V_{a^c}(x_1, \varepsilon_1)] \\
&\quad - [\Pr[g_a^*(X_a(x_1)) + \varepsilon_a^* > 0] - \Pr(g_a^*(X_a(x_1)) + \varepsilon_a^* > V_{a^c}^*(x_1, \varepsilon_1^*)]] \\
&\leqslant \Pr[g_a(X_a(x_1)) + \varepsilon_a > 0] - \Pr[g_a(X_a(x_1)) + \varepsilon_a > V_{a^c}(x_1, \varepsilon_1)] \\
&= \Pr[\mathrm{E}(\max(g_b(X_b) + \varepsilon_b, 0) \mid \varepsilon_1, x_1) \geqslant g_a(X_a(x_1)) + \varepsilon_a > 0].
\end{aligned}$$

I will now show that I can choose $x_1$ to set the final expression arbitrarily close to zero which leads to a contradiction.

Using Condition G3, by dominated convergence it is easy to show that for all $\varepsilon_1$,

$$\lim_{y \downarrow -S^{\mathrm{u}}_{\varepsilon_b}} \mathrm{E}(\max(y + \varepsilon_b, 0) \mid \varepsilon_1) = 0.$$

For any $\varepsilon_1$ and any $x_a \in \mathscr{X}_a^1$ we can then construct a sequence of random variables whose distribution is equivalent to the conditional distribution of $(\max(g_b(X_b) + \varepsilon_b, 0) \mid \varepsilon_1, x_{1,j})$ for a sequence of $x_{1,j} \in \mathscr{X}_1(x_a, y_j, c_j)$ where as $j \to \infty$, $y_j \downarrow -S^{\mathrm{u}}_{\varepsilon_b}$ and $c_j \to 1$. Applying the dominated convergence theorem to this sequence, one can show that $\mathrm{E}(\max(g_b(X_b) + \varepsilon_b, 0) \mid \varepsilon_1, x_{1j}) \to 0$. Thus we can find a $j$ large enough such that for $x_1 \in \mathscr{X}_1(x_a, y_j, c_j)$ we obtain a contradiction. $\square$

*Proof of Lemma 1.* By Assumption G4 we know that $g_a$ and $g_b$ are identified. Suppose that the lemma were false. Suppose that there exists a random vector $(\varepsilon_a^*, \varepsilon_b^*)$ whose distribution cannot be distinguished from that of the true random vector $(\varepsilon_a, \varepsilon_b)$. Then by definition of identification, for almost all $X$,

$$\Pr(g_a + \varepsilon_a > V_{a^c}(x_1, \varepsilon_1)) = \Pr(g_a + \varepsilon_a^* \leqslant V_{a^c}(x_1, \varepsilon_1^*)),$$

and

$$\begin{aligned}
&\Pr(g_a + \varepsilon_a \leqslant V_{a^c}(x_1, \varepsilon_1), g_b + \varepsilon_b > 0) \\
&\quad = \Pr(g_a + \varepsilon_a^* \leqslant V_{a^c}(x_1, \varepsilon_1^*), g_b + \varepsilon_b^* > 0),
\end{aligned} \tag{A.5}$$

but without loss of generality for some $\delta > 0$, since the joint distribution of $(\varepsilon_a, \varepsilon_b)$ is different from the joint distribution of $(\varepsilon_a^*, \varepsilon_b^*)$, there must be a set of $(g_a, g_b)$ with positive measure such that,

$$\Pr(\varepsilon_a^* \leqslant -g_a, -\varepsilon_b^* < g_b) - \Pr(\varepsilon_a \leqslant -g_a, -\varepsilon_b < g_b) > \delta. \tag{A.6}$$

But then for all members of this set and all $x_1 \in \mathrm{supp}(X_1)$ for which $g_a = g_a(X_a(x_1))$,

$$
\begin{aligned}
\delta &< \Pr(\varepsilon_a^* \leqslant -g_a, -\varepsilon_b^* < g_b) - \Pr(\varepsilon_a \leqslant -g_a, -\varepsilon_b < g_b) \\
&= \Pr[g_a + \varepsilon_a^* \leqslant 0, g_b + \varepsilon_b^* > 0] - \Pr[g_a + \varepsilon_a^* \leqslant V_{a^c}(x_1, \varepsilon_1^*), g_b + \varepsilon_b^* > 0] \\
&\quad - (\Pr[g_a + \varepsilon_a \leqslant 0, g_b + \varepsilon_b > 0] - \Pr[g_a + \varepsilon_a \leqslant V_{a^c}(x_1, \varepsilon_1), g_b + \varepsilon_b > 0]) \\
&\leqslant \Pr[g_a + \varepsilon_a \leqslant V_{a^c}(x_1, \varepsilon_1), g_b + \varepsilon_b > 0] - \Pr[g_a + \varepsilon_a \leqslant 0, g_b + \varepsilon_b > 0] \\
&= \Pr[V_{a^c}(x_1, \varepsilon_1) \geqslant g_a + \varepsilon_a > 0, g_b + \varepsilon_b > 0] \\
&\leqslant \Pr[V_{a^c}(x_1, \varepsilon_1) \geqslant g_a + \varepsilon_a > 0].
\end{aligned}
$$

Following exactly the last part of the proof of Theorem 1, I can show that there exists a set of $X_1$ with positive measure such that for $x_1$ in this set,

$$\Pr[V_{a^c}(x_1, \varepsilon_1) \geqslant g_a + \varepsilon_a > 0] < \delta.$$

but this is a contradiction, so the result must hold. □

*Proof of Theorem 2.* This follows trivially from Theorem 1 and Lemma 1 since the only unobservables in this case are $\varepsilon_a$ and $\varepsilon_b$. □

*Proof of Theorem 3.* I first show that the joint distribution of $(\varepsilon_a - v_b, v_b + \eta_b)$ is identified and then use this fact to show that both the distribution of $\eta_b$ and that the joint distribution of $(\varepsilon_a, v_b)$ are identified.

To see that the distribution of $(\varepsilon_a - v_b, v_b + \eta_b)$ is identified, recall that we normalized $V_c = 0$. This was arbitrary, we could have normalized $V_b = 0$. We can basically do that by redefining the model in the following manner:

$$\tilde{g}_a(X_a) = g_a(X_a) - \mathrm{E}(g_b(X_b) \mid X_1), \quad \tilde{g}_a(X_a) = g_a(X_a) - \mathrm{E}(g_b(X_b) \mid X_1), \tag{A.7}$$

$$\tilde{g}_c(X_c) = -g_b(X_b),$$

$$\tilde{g}_c(X_c) = -g_b(X_b), \tag{A.8}$$

$$\tilde{g}_b(X_a) = 0,$$

$$\tilde{g}_b(X_a) = 0. \tag{A.9}$$

We can then apply the lemma to the redefined model to obtain the desired result.

I now use the characteristic functions of these variables to complete the proof. I will make use of the notation $\phi_Y$ to denote the characteristic function of random variable $Y$ and $\phi_{Y_1 Y_2}$ to denote the characteristic function of the random vector $(Y_1, Y_2)$.

Suppose that there exist random variables $(\varepsilon_a^*, \eta_b^*, v_b^*)$ that generate the same choice probabilities as $(\varepsilon_a, \eta_b, v_b)$. First applying Lemma 1, we know that $\phi_{\varepsilon_a} = \phi_{\varepsilon_a^*}$. But $(\varepsilon_a - v_b, v_b + \eta_b)$ identified implies that,

$$\begin{aligned}
\phi_{\eta_b}(t) &= \frac{\mathrm{E}[\exp\{\mathrm{i}t(\varepsilon_a - v_b) + \mathrm{i}t(v_b + \eta_b)\}]}{\phi_{\varepsilon_a}(t)} \\
&= \frac{\mathrm{E}[\exp\{\mathrm{i}t(\varepsilon_a^* - v_b^*) + \mathrm{i}t(v_b^* + \eta_b^*)\}]}{\phi_{\varepsilon_a}(t)} \\
&= \phi_{\eta_b^*}(t),
\end{aligned}$$

so the distribution of $\eta_b$ is identified.

I can now show that the joint distribution of $v_b$ and $\varepsilon_a$ is identified since $\eta_b$ is independent of them and has a known distribution.

$$\begin{aligned}
\phi_{\varepsilon_a v_b}(t_1, t_2) &= \mathrm{E}[\exp\{\mathrm{i}t_1 \varepsilon_a + \mathrm{i}t_2 v_b\}] \\
&= \frac{\mathrm{E}[\exp\{\mathrm{i}t_1(\varepsilon_a - v_b) + \mathrm{i}(t_1 + t_2)(v_b + \eta_b)\}]}{\phi_{\eta_b}(t_1 + t_2)} \\
&= \frac{E[\exp\{\mathrm{i}t_1(\varepsilon_a^* - v_b) + \mathrm{i}(t_1 + t_2)(v_b^* + \eta_b^*)\}]}{\phi_{\eta_b}(t_1 + t_2)} \\
&= \phi_{\varepsilon_a^* v_b^*}(t_1, t_2).
\end{aligned}$$

The characteristic function and thus the distribution of $(\varepsilon_a, v_a)$ is identified and the full distribution of the unobservables is known. □

## References


Aakvik, A., Heckman, J., Vytlacil, E., 1999. Local instrumental variables and latent variable models for estimating treatment effects. Unpublished manuscript, University of Chicago.

Altonji, J., 1993. The demand for and return to education when education outcomes are uncertain. Journal of Labor Economics 11, 48–83.

Belzil, C., Hansen, J., 1997. Estimating the returns to education from a non-stationary dynamic programming model. Centre for Labour Market and Social Research Working Paper No. 97-06.

Buchinsky, M., Leslie, P., 1996. A dynamic model of education choices in the United States: learning from a cross-section. Unpublished manuscript, Brown University.

Cameron, S., Heckman, J., 1998. Life cycle schooling and dynamic selection bias: models and evidence for five cohorts of American males. Journal of Political Economy 106, 262–333.

Cameron, S., Taber, C., 1994. Assessing nonparametric maximum likelihood models of dynamic discrete choice. Unpublished manuscript.
Cameron, S., Taber, C., 1998. Borrowing constraints and the returns to schooling. Unpublished manuscript, Northwestern University.
Card, D., 1998. The causal effect of education on earnings. Center for Labor Economics, University of California at Berkeley Working Paper No. 2.
Chamberlain, G., 1986. Asymptotic efficiency in semi-parametric models with censoring. Journal of Econometrics 32, 189–218.
Comay, Y., Melnik, A., Pollatschek, M., 1973. The option value of education and the optimal path for investment in human capital. International Economic Review 14, 421–434.
Cosslett, S., 1983. Distribution-free maximum likelihood estimator of the binary choice model. Econometrica 51, 765–782.
Eckstein, Z., Wolpin, K., 1989. The specification and estimation of dynamic stochastic discrete choice models. Journal of Human Resources 24, 562–598.
Eckstein, Z., Wolpin, K., 1997. Youth employment and academic performance in high school. Unpublished manuscript, University of Pennsylvania.
Flinn, C., Heckman, J., 1982. New methods for analyzing structural models of labor force dynamics. Journal of Econometrics 18, 115–168.
Heckman, J., 1990. Varieties of selection bias. American Economic Review 80, 313–318.
Heckman, J., Honoré, B., 1990. The empirical content of the Roy model. Econometrica 58, 1121–1149.
Heckman, J., Singer, B., 1984. A method for minimizing the impact of distributional assumptions in economic models for duration data. Econometrica 52, 271–320.
Heckman, J., Vytlacil, E., 1999. Local instrumental variables and latent variable models for identifying and bounding treatment effects. Unpublished manuscript, University of Chicago.
Hotz, V.J., Miller, R., 1993. Conditional choice probabilities and the estimation of dynamic models. Review of Economic Studies 60, 497–529.
Ichimura, H., Taber, C., 1999. Estimation of treatment effect counter factuals under limited support conditions. Unpublished manuscript, Northwestern University.
Keane, M., Wolpin, K., 1997. The career decisions of young men. Journal of Political Economy 105, 473–522.
Manski, C., 1975. Maximum score estimation of the stochastic utility model of choice. Journal of Econometrics 3, 205–228.
Manski, C., 1988. Identification of binary response models. Journal of the American Statistical Association 83, 729–738.
Matzkin, R., 1990. Least concavity and the distribution-free estimation of nonparametric concave functions. Unpublished manuscript.
Matzkin, R., 1992. Nonparametric and distribution-free estimation of the binary threshold crossing and the binary choice models. Econometrica 60, 239–270.
Matzkin, R., 1993. Nonparametric identification and estimation of polychotomous choice models. Journal of Econometrics 58, 137–168.
Pakes, A., 1986. Patents as options: some estimates of the value of holding European patent stocks. Econometrica 54, 755–784.
Pakes, A., Simpson, M., 1989. Patent renewal data. In: Bailey, Winston (Eds.), Brookings Papers on Economic Activity. The Brookings Institute, Washington.
Roy, A.D., 1951. Some thoughts on the distribution of earnings. Oxford Economic Papers (New Series) 3, 135–146.
Rust, J., 1987. Optimal replacement of GMC bus engines: an empirical model of Harold zurcher. Econometrica 55, 999–1035.
Rust, J., 1994. Structural estimation of markov decision processes. In: Engle, R., McFadden, D. (Eds.), Handbook of Econometrics. North-Holland, Amsterdam.

Taber, C., 1996. Semiparametric identification and heterogeneity in discrete choice dynamic programming models. Unpublished manuscript, Northwestern University.

Taber, C., 1998. The rising college premium in the eighties: return to college or return to ability? Unpublished manuscript.

Thompson, T.S., 1989. Identification of semiparametric discrete choice models. Unpublished manuscript.

Weisbrod, B., 1962. Education and investment in human capital. Journal of Political Economy 70, 106–123.

Wolpin, K., 1987. Estimating a structural search model: the transition from school to work. Journal of Political Economy 55, 801–817.