TECHNICAL WORKING PAPER SERIES

DYNAMIC DISCRETE CHOICE AND DYNAMIC TREATMENT EFFECTS

James J. Heckman
Salvador Navarro

Technical Working Paper 316
http://www.nber.org/papers/T0316

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
October 2005

Dynamic Discrete Choice and Dynamic Treatment Effects
James J. Heckman and Salvador Navarro
NBER Technical Working Paper No. 316
October 2005
JEL No. C31

ABSTRACT

This paper considers semiparametric identification of structural dynamic discrete choice models and models for dynamic treatment effects. Time to treatment and counterfactual outcomes associated with treatment times are jointly analyzed. We examine the implicit assumptions of the dynamic treatment model using the structural model as a benchmark. For the structural model we show the gains from using cross equation restrictions connecting choices to associated measurements and outcomes. In the dynamic discrete choice model, we identify both subjective and objective outcomes, distinguishing *ex post* and *ex ante* outcomes. We show how to identify agent information sets.

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

Salvador Navarro
Department of Economics
University of Wisconsin - Madison
1180 Observatory Drive
Madison, WI 53706
snavarro@ssc.wisc.edu

1 Introduction

This paper presents econometric models for analyzing time to treatment and the consequences of the choice of a particular treatment time. Treatment may be a medical intervention, stopping schooling, opening a store, conducting an advertising campaign at a given date or renewing a patent. Associated with each treatment time, there can be multiple outcomes. They can include a vector of health status indicators and biomarkers; lifetime employment and earnings consequences of stopping at a particular grade of schooling; the sales revenue and profit generated from opening a store at a certain time; the revenues generated and market penetration gained from an advertising campaign; or the value of exercising an option at a given time. Our paper unites and contributes to the literatures on dynamic discrete choice and dynamic treatment effects. For both classes of models, we present semiparametric identification analyses.

The conventional treatment effect literature is static.[1] It ignores choice equations and only focuses on outcome equations.[2] We extend the literature on treatment effects to model choices of treatment times and the consequences of choice. We link the literature on treatment effects to the literature on precisely formulated structural dynamic discrete choice models generated from index models crossing thresholds. We show the value of precisely formulated economic models in extracting the information sets of agents, in providing model identification, in generating the standard treatment effects and in ruling out hard-to-interpret counterfactuals that can be generated from reduced form models.[3] With an articulated choice model in hand, it is possible to interpret, and relax, recent assumptions made in the treatment effect literature.

Our analysis of identification in dynamic discrete choice models is of interest in its own

[1] Robins (1989, 1997), Gill and Robins (2001) and Abbring and Van Den Berg (2003) are important contributions to the dynamic treatment effects literature.

[2] See Heckman (2006) and Heckman and Vytlacil (2006a).

[3] Aakvik, Heckman, and Vytlacil (2005), Heckman, Tobias, and Vytlacil (2001, 2003), Carneiro, Hansen, and Heckman (2001, 2003) and Heckman and Vytlacil (2005) show how standard treatment effects can be generated from structural models.

right. Rust (1994) provides a comprehensive survey of models in the field up to a decade ago and the field is burgeoning.[4] He shows that without additional restrictions, a class of infinite horizon dynamic discrete choice models for stationary environments is nonparametrically nonidentified.[5] His paper has fostered the widespread belief that dynamic discrete choice models are identified only by using arbitrary functional form and exclusion restrictions.[6] The entire dynamic discrete choice project thus appears to be without empirical content and the evidence from it at the whim of investigator choices about functional forms of estimating equations and application of *ad hoc* exclusion restrictions.

This paper establishes the semiparametric identifiability of a class of dynamic discrete choice models for stopping times and associated outcomes in which agents sequentially update the information on which they act. We also establish identifiability of a new class of reduced form duration models that generalize conventional discrete time duration models to produce frameworks with much richer time series properties for unobservables and general time-varying observables and patterns of duration dependence than conventional duration models. Our analysis of identification of discrete time duration models does not require conventional period-by-period exclusion restrictions. Instead, we rely on curvature restrictions across the index functions generating the durations that can be motivated by dynamic economic theory.[7]

The key to our ability to identify the structural model is that we supplement information on stopping times or time to treatment with additional information on measured consequences of choices of time to treatment as well as measurements. The current dynamic discrete choice literature focuses exclusively on the discrete choices. Economic theory generally imposes restrictions across transition and outcome equations. This information

[4] See Taber (2000); Magnac and Thesmar (2002) and Aguirregabiria (2004), among other important recent contributions.

[5] However, Rust's proof is for a stationary environment, infinite horizon, dynamic programming problem with recurrent states and does not use any information about concavity of utility functions or information connecting outcomes and choices.

[6] See for example the discussion in Magnac and Thesmar (2002).

[7] See Heckman and Honoré (1989, 1990) for examples of such an identification strategy in duration models and Roy models. See also Cameron and Heckman (1998).

provides identifying power only in fully articulated dynamic discrete choice models where choice equations are clearly delineated and are related to outcome equations, and not in reduced form analyses where the choice equation is left implicit. Our analysis demonstrates the power of economic theory in analyzing and interpreting models of treatment effects. With our structural framework, we can distinguish objective outcomes from subjective outcomes (valuations by the decision maker). Applying our analysis to health economics, we can identify the causal effects on health of a medical treatment as well as the associated subjective pain and suffering of a treatment regime for the patient. Attrition decisions also convey information about agent preferences about treatment.[8]

We do not rely on the assumption of conditional independence of unobservables, given observables, that is used throughout much of the dynamic discrete choice literature.[9] Similar assumptions underlie recent work on reduced form dynamic treatment effects in matching.[10] Our semiparametric analysis generalizes matching. In this paper, some of the variables that would produce conditional independence and would justify matching if they were observed are treated as unobserved match variables. They are integrated out and their distributions are identified.[11]

For specificity, throughout this paper we take as our principle example the choice of schooling and its consequences. Persons who start life in school may stop at different grades with consequences for their earnings, employment and other aspects of their socioeconomic trajectories. If each grade takes one period to complete, we can think of this model as a time to treatment model where the "treatment" is the grade at which a person "stops treatment" or drops out of school. Persons with different "treatment times" (attained levels of schooling) may have different lifetime employment and earnings outcomes while in school and afterward. Associated with each schooling attainment level (treatment time), may be

[8] See Heckman and Smith (1998). Use of participation data to infer preferences about outcomes is developed in Heckman (1974).

[9] See, *e.g.* Rust (1987), Manski (1993), Hotz and Miller (1993) and the papers cited in Rust (1994).

[10] See, *e.g.* Gill and Robins (2001) and Lechner and Miquel (2002).

[11] For estimates based on this idea see Carneiro, Hansen, and Heckman (2003), Aakvik, Heckman, and Vytlacil (2005), Cunha, Heckman, and Navarro (2005a,b,c,e), and Heckman and Navarro (2006).

measurements on IQ, genetic biomarkers and the like that may be used to proxy unobserved traits of the individuals being studied.

This paper proceeds in the following way. Section 2 presents our basic framework and establishes identification theorems for reduced form single spell duration models with general forms of duration dependence and heterogeneity. It is difficult to make some important economic distinctions within this model and the unrestricted model has some peculiar features that are difficult to interpret within a well posed economic model. Section 3 builds on the framework of Section 2 and develops identification conditions for a model of dynamic discrete choice and associated counterfactual outcomes with information updating and option values. Section 4 relates our analysis to previous work. Section 5 concludes. In a companion paper, Heckman and Navarro (2006), we apply our analysis to panel data on schooling choices and lifetime earnings to estimate both reduced form and structural models.

2 Semiparametric Duration Models and Counterfactuals

A basic building block for the analysis of this paper, of interest in its own right, is a semiparametric index model for dynamic discrete choices that extends conventional discrete time duration analysis. This framework can be used to approximate dynamic discrete choice models. The exact nature of the approximation is usually obscure, as is true of many models of treatment effects in economics and statistics. We allow for nonparametric duration dependence that can be generated by duration-specific regressors. We make explicit the unobservables that drive reduced form duration and heterogeneity dynamics. We separate out duration dependence from heterogeneity in a semiparametric framework more general than conventional discrete time duration models. We produce a new class of reduced form models for dynamic treatment effects by adjoining time-to-treatment outcomes to our duration model.

We first develop the time to treatment equation. In terms of our running example, the treatment time is the grade (age) at which a person stops schooling. The models we analyze throughout this paper are based on a latent variable for choice at time t by person i, $I_i(t) = \mu_t\left(Z_i(t), \eta_i(t)\right)$, where the $Z_i(t)$ are observables and $\eta_i(t)$ are unobservables from the point of view of the econometrician. In Section 3, we derive $I_i(t)$ from a specific economic model. Treatments at different times may have different outcome consequences which we model after analyzing the time to treatment equation. Define $D_i(t)$ as an indicator of receipt of treatment at date t for individual i. Treatment is taken the first time $I_i(t)$ becomes positive. Thus $D_i(t) = \mathbf{1}\left[I_i(t) > 0, I_i(t-1) \leq 0, I_i(t-2) \leq 0, \ldots\right]$ where the indicator function $\mathbf{1}[\cdot]$ takes the value of 1 if the term inside the braces is true.[12] We derive conditions for identifying a model with general forms of duration dependence in the time to treatment equation. To simplify notation, we drop the "i" subscript throughout the paper. In discussing identification, we assume access to panel data on individuals with observations statistically independent across persons, but potentially dependent across time for the same person.

2.1 Single Spell Duration Model

Individuals are assumed to start spells in a given (exogenously determined) state and to exit the state at the beginning of time period $T = t$.[13] In our schooling example, an individual starts school and drops out in period T. T is thus a random variable representing total completed spell length. It can also be interpreted as time to treatment (*i.e.*, the agent waits in the no treatment state $t-1$ periods and exits into treatment at the beginning of period $T = t$).[14] Let $D(t) = 1$ if the individual exits at time t and $D(t) = 0$ otherwise. In our

[12] This framework captures the essential feature of any stopping time model. For example, in a search model with one wage offer per period, $I_i(t)$ is the gap between market wages and reservation wages at time t. See, *e.g.* Flinn and Heckman (1982). This framework can also approximate the explicit dynamic discrete choice model analyzed in Section 3.

[13] Thus we abstract from the initial-conditions problem discussed in Heckman (1981b).

[14] $T = t$ designates either completion of a treatment regime at t or else the date at which treatment is received.

schooling example where each year of school takes one period to complete, t is the number of years of completed schooling for people who start in school.[15] Treatment at t consists of dropping out at the beginning of period t. The event $D(t-1) = 0$ signifies that an individual remains in the no treatment state at $t-1$. We impose an exogenously specified initial condition $D(0) = 0$. In a schooling example, $\bar{T}$ is the highest possible grade that can be completed and $D(0) = 0$ means that everyone starts with zero years of schooling.

In an analysis of drug treatments, $T = t$ is the discrete time period in the course of an illness at the beginning of which the drug is administered. There may be a maximum duration of the illness $\bar{T}$ beyond which treatment cannot be administered. It is possible in this example that $D(0) = 0, \ldots, D(\bar{T}) = 0$, so that a patient never receives treatment. In the schooling example, "treatment" is not schooling, but rather dropping out of schooling. In this case, if there is an upper limit $\bar{T}$ to the number of years of schooling, if $D(0) = 0, \ldots, D(\bar{T}-1) = 0$, then $D(\bar{T}) = 1$. Our analysis applies to both cases, but we focus on the schooling example because it links the analysis of this section to the analysis of Section 3.

In the context of this model, there is no meaningful event corresponding to the outcome $D(t) = 0$ and $D(t-1) = 1$, so the $D(t)$ have a natural sequential structure: $(D(0), D(1), \ldots, D(t)) = (0, 0, \ldots, 1)$. For a given stopping time t, we denote by D^t the truncated sequence consisting of the first $t+1$ elements (from 0 to t) of D. In the course of our discussion, we will make use of the random variables $D(t)$ and D^t for fixed t, $t = 1, \ldots, \bar{T}$. By abuse of notation, we will designate by $d(t)$ and d^t values that these two random variables can assume. Thus, $d(t)$ can be zero or one and d^t is a sequence of $t+1$ elements consisting of a nonempty subsequence of zeros followed by a (possibly empty) subsequence of ones. For a sequence of all zeros, we will write $D^t = (0)$ and $d^t = (0)$ regardless of the length of these subsequences. Let $Z(t) = z(t)$ denote regressors determining transitions from time $t-1$ to time t. Let $\bar{T}$ $(< \infty)$ be the upper limit on the time the agent being studied can be at risk for a treatment.

[15] We assume that once out of school a person does not attend again. Alternatively, we use years attended rather than grade completed as the measure of schooling.

Our duration model arises from the threshold-crossing behavior of a sequence of underlying latent indices:

$$\left.\begin{array}{rcl} D(t) & = & \mathbf{1}\left[I(t) \geq 0\right] \\ I(t) & = & Z(t)\gamma_t - \eta(t) \end{array}\right\} \text{ if } D^{t-1} = (0),\ t = 1, \ldots, \bar{T}, \tag{1}$$

where $\mu_t\left(Z(t), \eta(t)\right) = Z(t)\gamma_t - \eta(t)$. The $D(t)$ outcome is observed only if $D(t-1) = 0$, which is equivalent to $D^{t-1} = (0)$. The $Z(t)$ are regressors that enter the index at period t. The $Z(t)$ can include expectations of future outcomes given current information in the case of models with forward-looking behavior. To identify period t parameters from period t outcomes, one must condition on all past outcomes and control for any selection effects. The assumption of linearity of the index in $Z(t)$ is not critical to our analysis, and this assumption can be relaxed following arguments in Matzkin (1992, 1993, 1994). Appendix B presents the class of nonparametric functions identified by Matzkin. We call them Matzkin functions. Using Matzkin (2003), we can also relax the separability assumption, but we do not do so in this paper.

Let $Z = \left(Z(1), \ldots, Z(\bar{T})\right)$, and let $\eta = (\eta(1), \ldots, \eta(\bar{T}))$.[16] We assume that Z is statistically independent of η. Let $\gamma = (\gamma_1, \ldots, \gamma_{\bar{T}})$. Depending on the values assumed by γ_t, we can generate very general forms of duration dependence that depend on the values assumed by the $Z(t)$. We thus allow for period-specific effects of regressors on the latent indices generating choices.

This model is the reduced form of a general dynamic discrete choice model. Like many reduced form models, the link to choice theory is not clearly specified. It is not a conventional multinomial choice model in a static (perfect certainty) setting with associated outcomes. As

[16] A special case of the general model arises when $\eta(t)$ has a factor model representation,

$$\eta(t) = \alpha_t \theta + \varepsilon(t),\ t = 1, \ldots, \bar{T} \text{ where } \alpha_1 = 1,$$

where we assume that $\varepsilon(t) \perp\!\!\!\perp \varepsilon(t')$, for $t \neq t'$, that $\varepsilon(t) \perp\!\!\!\perp \theta$, where "$\perp\!\!\!\perp$" denotes statistical independence, and that $(\theta, \varepsilon(1), \ldots, \varepsilon(\bar{T}))$ is jointly independent of Z. Setting $\alpha_t = 1$ for all t generates the conventional permanent-transitory model.

a point of reference, we present such a model in Appendix C and consider its identifiability. We analyze the model based on equation (1) because it extends conventional discrete time duration analysis and because our analysis of identification in this simple setting produces results that are useful for securing identification in the more explicit structural model of Section 3.

2.2 Identification of Duration Models with General Error Structures and Duration Dependence

We first establish semiparametric identification of the model of equation (1). We assume access to a large sample of i.i.d. (D, Z) observations. Let $Z^t = (Z(1), \ldots, Z(t))$, $\gamma^t = (\gamma_1, \ldots, \gamma_t)$. We can nonparametrically identify the conditional probability $\Pr(D(t) = d(t) | Z^t, D^{t-1} = d^{t-1})$ *a.e.* $F_{Z^t|D^{t-1}=d^{t-1}}$ where $F_{Z^t|D^{t-1}=d^{t-1}}$ is the distribution of Z conditional on previous choices. We assume that $(\gamma, F_\eta) \in \Gamma \times \mathcal{H}$, where $\Gamma \times \mathcal{H}$ is the parameter space. Our goal is to establish conditions under which knowledge of $\Pr(D(t) = d(t) | Z, D^{t-1} = d^{t-1})$ *a.e.* $F_{Z|D^{t-1}=d^{t-1}}$ allows us to identify a unique element of $\Gamma \times \mathcal{H}$. We define identification in a standard way.

Definition 1. *Let $P_{\gamma^t, F_{\eta^t}}(D(t) = 1 | Z^t = z^t, D^{t-1} = d^{t-1})$ be the probability of observing the choice $D(t) = 1$ conditional on observables $Z^t = z^t$ and past choices $D^{t-1} = d^{t-1}$ under model (1) when the parameter values are given by (γ^t, F_{η^t}). Let $\Gamma \times \mathcal{H}$ be the space of permissible parameter values. We say that $(\gamma^t, F_{\eta^t}) \in \Gamma \times \mathcal{H}$ is identified iff for all $\left(\gamma^{*t}, F^*_{\eta^t}\right) \in \Gamma \times \mathcal{H} \backslash (\gamma^t, F_{\eta^t})$, there exists a sequence of past choices, d^{t-1}, $Pr(D^{t-1} = d^{t-1}) > 0$, such that*

$$\Pr{}_{Z^t|D^{t-1}=d^{t-1}} \left\{ P_{\gamma^t, F_{\eta^t}}(D(t) = 1 | Z^t, D^{t-1} = d^{t-1}) \neq P_{\gamma^{*t}, F^*_{\eta^t}}(D(t) = 1 | Z^t, D^{t-1} = d^{t-1}) \right\} > 0.$$ [17]

To secure identification of all of the models in this paper, we follow an identification-in-

[17] Alternatively, we could define identification in terms of the joint distribution of $D(t)$ and Z given $D^{t-1} = d^{t-1}$ rather than in terms of the conditional distribution of $D(t)$.

the-limit strategy that allows us to recover the (γ^t, F_{η^t}) by conditioning on large values of the indices of the preceding choices. This identification strategy is widely used in the analysis of discrete choice.[18]

We now establish sufficient conditions for the identification of model (1).

Theorem 1. *For the model defined by equation (1), assume the following conditions:*

(i) $\eta^t \equiv (\eta(1), \ldots, \eta(t))$ *is statistically independent of* $Z^t = (Z(1), \ldots, Z(t))$, $t = 1, \ldots, \bar{T}$,

(ii) η^t *is a continuous random variable*[19] *on* $\mathbb{R}^t$ *with support* $\prod_{j=1}^{t} \big(\underline{\eta}(j), \bar{\eta}(j)\big)$, *where* $-\infty \leq \underline{\eta}(j) < \bar{\eta}(j) \leq +\infty$ *for all* $j = 1, \ldots, \bar{T}$, *and the joint distribution does not depend on* γ^t,

(iii) (Full Rank of $Z(t)$*) For all* $j = 1, \ldots, t$, $Z(t)$ *is a* K_t*-dimensional random variable. There exists no proper linear subspace of* $\mathbb{R}^{K_t}$ *having probability 1 under* $F_{Z(t)}$. *There exists a* $\check{g}^t = (\check{g}_1, \ldots, \check{g}_{t-1})$ *such that for almost every* $g^t = (g_1, \ldots, g_{t-1}) \in \prod_{j=1}^{t-1} \big(\underline{\eta}(j), \bar{\eta}(j)\big)$ *with* $g^t \geq \check{g}^t$ *(componentwise), there exists no proper linear subspace of* $\mathbb{R}^{K_t}$ *having probability 1 under* $F_{Z(t)|Z(1)\gamma_1 \geq g_1, \ldots, Z(t-1)\gamma_{t-1} \geq g_{t-1}}$.

(iv) (Inclusion of Supports) Supp $\big(Z(t)\gamma_t | Z(1)\gamma_1 = g_1, \ldots, Z(t-1)\gamma_{t-1} = g_{t-1}\big) \supseteq \big(\underline{\eta}(t), \bar{\eta}(t)\big)$ *for almost every* $(g_1, \ldots, g_{t-1}) \in \prod_{i=1}^{t-1} \big(\underline{\eta}(t), \bar{\eta}(t)\big)$, *for* $t = 1, \ldots, \bar{T}$, *where the boundary points* $\big\{\underline{\eta}(t), \bar{\eta}(t) : t = 1, \ldots, \bar{T}\big\}$ *are not functions of* γ_t *for* $t = 1, \ldots, \bar{T}$, *where "Supp" means support. The supports can be unbounded.*

Then F_{η^t} *and* (γ^t) *are identified given location and scale normalizations,* $t = 1, \ldots, \bar{T}$.

Proof. See Appendix C. ■

[18] See, *e.g.* Manski (1988), Heckman (1990), Heckman and Honoré (1989, 1990), Matzkin (1992, 1993), Taber (2000), and Carneiro, Hansen, and Heckman (2003). A version of the strategy of this proof was first used in psychology where agent choice sets are eliminated by experimenter manipulation. The limit set argument effectively uses regressors to reduce the choice set confronting agents. See Falmagne (1985) or Thurstone (1959).

[19] We say a random variable is "continuous" if it is absolutely continuous with respect to Lebesgue measure.

Assumption (*iii*) is used to guarantee full rank of the model in limit sets where the probability of events becomes arbitrarily small. In place of assumption (*iv*), one can work with a more general index $\Psi(t, Z(t))$ to replace $Z(t)\gamma_t$ and identify it over the relevant support, which can be bounded if $\Psi(t, Z(t))$ belongs to the Matzkin class of functions presented in Appendix A. We use this more general nonseparable model in Theorems 2 and 3 and Corollary 2, and a fully nonseparable choice model in Section 3 below. Independence assumption (i) is strong. A more general version of Theorem 1 can be proved using the analysis of Lewbel (2000).[20]

The assumptions of Theorem 1 will be satisfied if there are transition-specific exclusion restrictions for Z with the required properties. In models with many periods, this may be a demanding requirement. Very often, the Z variables are time invariant and so cannot be used as exclusion restrictions. The following corollary tells us that the model can be identified even if there are no conventional exclusion restrictions and the $Z(t)$ are the *same* across all time periods if sufficient structure is placed on how the γ_t vary with t. Variations in the values of γ_t across time periods arise naturally in finite horizon dynamic discrete choice models where a shrinking horizon produces different effects of the same variable in different periods. For example, in the analysis of a search model by Wolpin (1987), the value function depends on time and the derived decision rules weight the same invariant characteristics differently in different periods. In a schooling model, parental background and resources may affect education continuation decisions differently at different stages of the schooling decision. The model generating equation (1) can be semiparametrically identified without transition-specific exclusions if the duration dependence is sufficiently general.

Corollary 1. *For the model defined by equation (1), suppose in addition to the conditions (i)–(iv) of Theorem 1 that*

(v) In condition (iii), $Z(t) = Z$ for all t where Z is a K-dimensional random variable.

[20] Magnac and Maurin (2005) show how to use the Lewbel regressor to bypass identification at infinity arguments. The conditions required for application of Lewbel's theorem and its extensions are not easily satisfied. See Theorem 1′ and its proof at our website, http://jenni.uchicago.edu/dyn-trmt-eff, where an extension of Theorem 1 using the Lewbel special regressor is presented.

Thus the same regressors are assumed to appear in all transitions. We define Z so that the first T^* coordinates of Z are continuous random variables ($T^* \leq K$). The support of the first T^* coordinates of Z is $\prod_{i=1}^{T^*}(-\infty, \infty)$.

(vi) *$\gamma_1, \ldots, \gamma_{T^*}$, the coefficients associated with the Z for the first T^* periods of the spell, are linearly independent. Denote the i^{th} component of t by γ_t^i, $(i = 1, \ldots, K)$. The first T^* coordinates of the γ_t, are non-zero for all $t = 1, \ldots, T^*$.*

Under these conditions, assumptions (iii) and (iv) of Theorem 1 are satisfied with $\underline{\eta}(i) = -\infty, \bar{\eta}(i) = \infty$. Given assumptions (i) and (ii) of Theorem 1 and assumptions (v)–(vi) just given, F_{η^t} where η^t and $\gamma^t, t = 1, \ldots, \bar{T}^*$ are identified up to scale and location normalizations, where $\gamma^t = (\gamma_1, \ldots, \gamma_t)$ is to be distinguished from the i^{th} component of γ_t denoted γ_t^i.

Proof. See Appendix C. ■

If $T^* < \bar{T}$, full identification of the model is not possible without additional information. Observe that the number of periods where the γ_t are identified and joint distribution of the $\eta(1), \ldots, \eta(t)$ is identified depends crucially on the number of continuous regressors. If there are fewer continuous regressors (T^*) than time periods, ($\bar{T}$), the most we can identify are the parameters $\gamma_1, \ldots, \gamma_t$ and the joint distribution F_{η^t} for $t = T^*$.

Conditions *(v)* and *(vi)* are sufficient conditions for producing measurable separability or "variation freeness" among the indices.[21] Using the Matzkin class of functions described in Appendix B, we can extend this analysis to a general model that is nonseparable in (Z, t) but separable in $\eta(t)$. In Section 3 we prove a result analogous to Corollary 1 for a structural model using the general representation for a more general choice function that is fully nonseparable in all of its arguments. Theorem 1 and its Corollary provide a specific example of functions that satisfy the more general, "measurable separability" condition that is the fundamental principle underlying identification in this class of models.[22]

[21] See Florens, Mouchart, and Rolin (1990, pp. 189–200) for a precise definition of measurable separability. This concept clarifies the notion of "variation free" variables.

[22] See Theorems 5 and 7 in Section 3.

Theorem 1 and Corollary 1 have important consequences. The $Z(t)\gamma_t$, $t = 1, \ldots, \bar{T}$ (or more generally the $\Psi(t, Z)$) can be interpreted as duration dependence parameters that are modified by the $Z(t)$ and that vary across the spell in a more general way than is permitted in mixed proportional hazards (MPH), generalized accelerated failure time ($GAFT$) models or standard discrete time hazard models.[23] Duration dependence in conventional specifications of duration models is usually generated by variation in model intercepts. We allow the regressors to interact with the duration dependence parameters. The "heterogeneity" distribution F_η is identified for a general model. No special "permanent-transitory" structure is required for the unobservables although that specification is traditional in duration analysis. Our explicit treatment of the stochastic structure of the duration model is what allows us to link in a general way the unobservables generating the duration model to the unobservables generating the outcome equations that are introduced in the next section. Such an explicit link is not currently available in the literature on continuous time duration models for treatment effects, and is useful for modelling selection effects in outcomes across different treatment times. Our outcomes can be both discrete and continuous and are not restricted to be durations.

Under the rank condition on the γ_t, no period-specific exclusion conditions are required on the Z. Abbring and Van Den Berg (2003) note that period-specific exclusions are not natural in reduced form duration models designed to approximate forward-looking life cycle models. Agents make current decisions in light of their forecasts of future constraints and opportunities, and if they forecast some components well, and they affect current decisions, then they are in $Z(t)$ in period t. The rank condition of Corollary 1 and its extension in Section 3 are of great value in establishing identification without such exclusions. We now adjoin a system of counterfactual outcomes to our model of time to treatment to produce a model for dynamic counterfactuals.

[23] See Ridder (1990) for a discussion of these models.

2.3 Reduced Form Dynamic Treatment Effects

This section develops a reduced form approach to generating dynamic counterfactuals. We apply and extend the analysis of Carneiro, Hansen, and Heckman (2003), henceforth CHH, to generate *ex post* potential outcomes and their relationship with the time to treatment indices $I(t)$ analyzed in the preceding subsection. With reduced form models, it is difficult to impose restrictions from economic theory or to make distinctions between *ex ante* and *ex post* outcomes. In the structural model developed in Section 3, these and other distinctions can be made easily.

Associated with each treatment time t is a vector of outcomes $Y(t, X, U(t)), t = 1, \ldots, \bar{T}$. Elements of this vector are outcome states associated with stopping (receiving treatment) at the beginning of period t. For stopping times t' different from t, $Y(t', X, U(t')), t' \neq t$, $t' = 1, \ldots, \bar{T}$ are counterfactuals. They depend on observables, X, and unobservables, $U(t)$, where the observability distinction is made from the point of view of the econometrician. The X may be t specific but for the sake of notational simplicity we use the simple X notation. The outcome variables are not necessarily what the agent thinks will happen when he or she stops at any particular date t, but rather what actually happens. The reduced form approach presented in this section is not sufficiently rich to precisely capture the notion that agents revise their anticipations of $Y(t, X, U(t)), t = 1, \ldots, \bar{T}$ as they acquire information over time. This notion is systematically developed using the structural model of Section 3.

The treatment "times" may be stages that are not necessarily connected with real times. Thus in the analysis of section 3, "t" is a schooling level. The correspondence between stages and times is exact if each stage takes one period to complete. Our notation is more flexible, and time and periods can be defined more generally. Our notation in this section accommodates both cases.

It is possible to think of $Y(t, X, U(t))$ as a vector of outcomes with components revealed

at each age, $a = 1, \ldots, \bar{A}$:

$$Y\left(t, X, U\left(t\right)\right) = \left(Y\left(1, t, X, U\left(1, t\right)\right), \ldots, Y\left(a, t, X, U\left(a, t\right)\right), \ldots, Y\left(\bar{A}, t, X, U\left(\bar{A}, t\right)\right)\right),$$

where we define $U\left(t\right) = \left(U\left(1, t\right), \ldots, U\left(a, t\right), \ldots, U(\bar{A}, t)\right)$. The X may also have age and t specific subvectors $\left(a = 1, \ldots, \bar{A};\ t = 1, \ldots, \bar{T}\right)$. Henceforth, whenever we have random variables with multiple arguments $R_0(t, Q_0, \ldots)$ or $R_1(a, t, Q_0, \ldots)$ where the argument list begins with time t or both age a and time t (perhaps followed by other arguments $Q_0, \ldots$), we will make use of several condensed notations: (a) dropping the first argument as we collect the components into vectors $R_0(Q_0, \ldots)$ or $R_1(t, Q_0, \ldots)$ of length $\bar{T}$ or $\bar{A}$, respectively, and (b) going further in the case of R_1, dropping the t argument as we collect the vectors $R_1(t, Q_0, \ldots)$ into a single $\bar{T} \times \bar{A}$ array $R_1(Q_0, \ldots)$.

This notation is sufficiently rich to represent the life cycle of outcomes for persons who receive treatment at t. Thus, in a schooling example, the components of this vector may include life cycle earnings, employment, and the like associated with a person with characteristics X, $U\left(t\right)$, $t = 1, \ldots, \bar{T}$, who completes t years of schooling and then forever ceases schooling. It could include earnings while in school at some level for persons who will eventually attain further schooling as well as post school earnings. Measuring a and t in the same units, we initialize the process by assuming that $t = 0$ and $a = 0$.

The $Y\left(a, t, X, U\left(a, t\right)\right)$ for $a < t$ are outcomes realized while the person is in school at age a (t is the time the person will leave school; a is the current age) and before "treatment" (stopping schooling) has occurred. When $a \geq t$, these are post-school outcomes for treatment with t years of schooling. In this case, $a - t$ is years of post-school experience. In the case of a drug trial, the $Y\left(a, t, X, U\left(a, t\right)\right)$ for $a < t$ are measurements observed before the drug is taken at t and if $a \geq t$, they are the post-treatment measurements.

Following CHH, the variables in $Y\left(a, t, X, U\left(a, t\right)\right)$ may include discrete, continuous or mixed discrete-continuous components. For the discrete or mixed discrete-continuous cases,

we assume that latent continuous variables cross thresholds to generate the discrete components. Durations can be generated by latent index models associated with each outcome crossing thresholds analogous to the model presented in equation (1). In this framework, we can model the effect of attaining t years of schooling on durations of unemployment or durations of employment.

Each treatment time can have its own age path of *ex post* outcomes even after correcting for selection effects by controlling for observed and unobserved determinants of outcomes apart from treatment time, and thus controlling for selection effects. In addition, paths prior to treatment may be different for different treatment times. Thus we can allow earnings at age a for people who receive treatment at some future time t' to differ from earnings at age a for people who receive treatment at some future time t'', $\min(t', t'') > a$ even after controlling for $U(t)$ and X.[24]

In a model with uncertainty, agents act on and value *ex ante* outcomes. The model developed in Section 3 distinguishes *ex ante* from *ex post* outcomes. The model developed in this section cannot because, within it, it is difficult to specify the information sets on which agents act or the mechanism by which agents forecast and act on $Y(t, X, U(t))$ when they are making choices.

One justification for not making an *ex ante* – *ex post* distinction is that the agents being modeled operate under perfect foresight even though econometricians do not observe all of the information available to the agents. In this framework, the $U(t), t = 1, \ldots, \bar{T}$, are an ingredient of the econometric model that accounts for the asymmetry of information between the agent and the econometrician studying the agent.

Without imposing assumptions about the functional structure of the outcome equations, we cannot nonparametrically identify counterfactual outcome states $Y(t, X, U(t))$ that have never been observed. Thus, in the schooling example, we assume that we observe life cycle

[24] Thus we do not need to impose the "no anticipations" assumption of Abbring and Van Den Berg (2003). However, it arises naturally in a fully specified structural model as we note in Section 3 and in Abbring and Heckman (2006).

outcomes for some persons for each stopping time (level of final grade completion) and our notation reflects this.[25] However, we do not observe $Y(t, X, U(t))$ for all t for anyone. A person can have only one stopping time (one completed schooling level). This observational limitation creates the "fundamental problem of causal inference."[26]

In addition to this problem, there is the standard selection problem that the $Y(t, X, U(t))$ are only observed for persons who stop at t and not for a random sample of the population. The selected distribution may not accurately characterize the population distribution of $Y(t, X, U(t))$ for persons selected at random. Note also that without further structure, we can only identify treatment responses within a given policy environment. In another policy environment where the rules governing selection into treatment and/or the outcomes from treatment may be different, the same time to treatment may be associated with entirely different responses.[27] We now turn to an analysis of identification.

2.4 Identification of Outcome and Treatment Time Distributions

We assume access to data on $(T, Y(T, X, U(T)), X, Z)$ for persons for whom $T = t$, $X = x$, $Z = z$ where T is the stopping time, X are the variables determining outcomes and Z are the variables determining choices. We also assume that we know $\Pr(T = t \mid Z = z)$ for $t = 1, \ldots, \bar{T}$. We assume independence of all outcomes across persons. Appendix D presents a general analysis of identification for vector valued $Y(T, X, U(T))$. In the text, we consider three special cases: (a) outcomes are scalar continuous variables (*e.g.* present value of earnings for a schooling example), (b) outcomes are discrete but vector valued (*e.g.* employment at each age) and (c) outcomes are durations (*e.g.* spells of unemployment). The first case is developed further in Section 3. The third case is a discrete time analogue of the model for counterfactual duration distributions analyzed by Abbring and Van Den Berg

[25] In practice we can only observe a portion of the life cycle after treatment. See the discussion on pooling data in Cunha, Heckman, and Navarro (2005e) to replace missing life cycle data. See Heckman and Vytlacil (2005) for analyses of how to construct never-observed counterfactuals.

[26] See Holland (1986) or Gill and Robins (2001).

[27] This is the problem of general equilibrium effects. See Heckman, Lochner, and Taber (1998), Heckman, LaLonde, and Smith (1999) or Abbring and Van Den Berg (2003) for discussion of this problem.

(2003).

We first consider the analysis of continuous outcomes. Our results generalize the analysis of Heckman and Honoré (1990), Heckman (1990) and CHH by considering choices generated by a stopping time model. To simplify the notation in this section, we assume that the scalar outcome associated with stopping at time t can be written as $Y(t) = \mu(t, X) + U(t)$, where $Y(t)$ is shorthand for $Y(t, X, U(t))$. We observe $Y(t)$ only if $D^{t-1} = (0)$, $D(t) = 1$ where the $D(t)$ are generated by a more general version of the index for time to treatment than was used in the analysis of Theorem 1 and Corollary 1. We replace $Z_t\gamma_t$ by $\Psi(t, Z)$ and write $I(t) = \Psi(t, Z) - \eta(t)$. We assume that the $\Psi(t, Z)$ belong to the Matzkin class of functions described in Appendix B. In the following, we will make use of the condensed representations I, $\Psi(Z)$, η, Y, $\mu(X)$ and U as described in Section 2.3.

We permit general stochastic dependence within the components of U, within the components of η and across the two vectors. We assume that (X, Z) are independent of (U, η). Each component of (U, η) has a zero mean. The joint distribution of (U, η) is assumed to be absolutely continuous. Recall that we allow the $X(t)$ to vary period by period. To simplify notation, we simply condition on the entire vector of the X.

With "sufficient variation" in the components of $\Psi(Z)$, we can identify $\mu(t, X)$, $[\Psi(1, Z(1)), \ldots, \Psi(t, Z(t))]$ and the joint distribution of $U(t)$ and η^t. This enables us to identify average treatment effects across all stopping times, since we can extract $E(Y(t) - Y(t') \mid X = x)$ from the marginal distributions of $Y(t)$, $t = 1, \ldots, \bar{T}$.

Theorem 2. *Assume data on $(Y(t), X, Z)$ given $T = t$ from a random sample across persons. We also observe (T, Z) from a random sample and we assume that the T are not censored. Write $\eta^t = (\eta(1), \ldots, \eta(t))$ and $\Psi^t(Z) = (\Psi(1, Z(1)), \ldots, \Psi(t, Z(t)))$. The $\Psi^t(Z)$ are elements of the Matzkin class of functions. Assume that*

(i) $(U(t), \eta^t)$ are continuous random variables with zero means, finite variances and with support $Supp(U(t)) \times Supp(\eta^t)$ with upper and lower limits $\left(\bar{U}(t), \bar{\eta}^t\right)$ and $\left(\underline{U}(t), \underline{\eta}^t\right)$ respectively, $t = 1, \ldots, \bar{T}$. These conditions hold for each component of each subvector.

The joint system is thus measurably separable for each component with respect to every other component.

(ii) $(U(t), \eta^t) \perp\!\!\!\perp (X, Z)$, $t = 1, \ldots, \bar{T}$ *(independence).*

(iii) $Supp\left(\mu\left(t, X\right), \Psi^t\left(Z\right)\right) = Supp\left(\mu\left(t, X\right)\right) \times \prod_{j=1}^{t} Supp\left(\Psi(j, Z\left(j\right))\right)$, $t = 1, \ldots, \bar{T}$.

(iv) $Supp\left(\Psi^t\left(Z\right)\right) \supseteq Supp\left(\eta^t\right)$

Then we can identify $\mu\left(t, X\right), \Psi^t\left(Z\right), F_{\eta^t, U(t)}$, $t = 1, \ldots, \bar{T}$, *up to scale if the Matzkin class is specified up to scale, and are exactly identified if a specific normalization is used.*

Proof. From data on $Y\left(t\right)$, X and Z for $D\left(t\right) = 1$, $D^{t-1} = (0)$, and from data on stopping times for the entire sample, we can identify for each $X = x$ and $Z = z$ the left hand side of the equation

$$\Pr\left(Y\left(t\right) < y\left(t\right) \mid D\left(t\right) = 1, D^{t-1} = (0), X = x, Z = z\right)$$
$$\times \Pr\left(D\left(t\right) = 1, D^{t-1} = (0) \mid X = x, Z = z\right)$$

$$= \int_{\underline{U}(t)}^{y(t)-\mu(t,x)} \int_{\underline{\eta}(t)}^{\Psi(t,z)} \int_{\Psi(t-1,z(t-1))}^{\bar{\eta}(t-1)} \cdots \int_{\Psi(1,z(1))}^{\bar{\eta}(1)} f_{U(t),\eta^t}\left(u, \eta(1), \ldots, \eta(t)\right) d\eta(1) \cdots d\eta(t)\, du. \quad (2)$$

$D\left(0\right) = 0$ is fixed exogenously outside of the model.

Under assumption *(iv)*, for all $x \in Supp\left(X\right)$ we can vary the values of Z and obtain a limit set $\mathcal{Z}$ such that $\lim_{Z \to \mathcal{Z}} \Pr\left(D\left(t\right) = 1, D^{t-1} = (0) \mid X = x, Z = z\right) = 1$. Thus we can identify the distribution of $U\left(t\right)$, $t = 1, \ldots, \bar{T}$, free of selection bias. From this argument, we can identify the $\mu\left(t, X\right)$. (We recover the intercepts through the assumption $E\left(U(t)\right) = 0$.) Condition *(iv)* allows us to generalize Theorem 1 by allowing for a more general specification of the index functions belonging to the Matzkin class. Using her analysis we can recover the $\Psi\left(t, Z\right)$. From knowledge of $y\left(t\right)$ and $\mu\left(t, X\right)$, $\Psi^t(Z)$, and from condition *(iii)*, we can vary $y\left(t\right) - \mu\left(t, X\right)$, $\Psi^t(Z)$ freely and trace out the joint distribution of $\left(U\left(t\right), \eta^t\right)$. Under

the assumptions of the theorem, we can do this for all $t = 1, \ldots, \bar{T}$. If we use the Matzkin conditions for functions up to scale, we identify the $\Psi^t(Z)$ up to scale and the distributions of the unobservables up to scale $F_{\eta^t, U(t)}$, $t = 1, \ldots, \bar{T}$. ■

Theorem 2 does not identify the joint distribution of $Y(1), \ldots, Y\left(\bar{T}\right)$ because we observe only one of these outcomes for any person. Observe that we do not require exclusion restrictions in the arguments of the choice of treatment equation to identify the counterfactuals. We require independent variation of arguments ("measurable separability") which might be achieved by exclusion conditions but can be obtained by other functional restrictions as in the proof of Corollary 1. Observe further that we can identify the $\mu(t, X)$ (up to constants) without the limit set argument. From the expression for (2), for each fixed $Z = z$ and $\Pr(D(t) = 1, D^{t-1} = (0) \mid X = x, Z = z) = p$, we can vary $y(t)$ and trace out $\mu(t, X)$ within each p set (see Heckman, 1990; Heckman and Smith, 1998, and CHH). Thus we can identify certain features of the model without using the limit set argument.

The proof of Theorem 2 can easily be extended to cover the case of vector $Y(t, X, U(t))$ where each component is a continuous random variable. See Theorem D.1 in Appendix D. There we allow for age-specific outcomes $Y(a, t, X, U(a, t))$, $a = 1, \ldots, \bar{A}$ where Y can be a vector of outcomes. In particular, we can identify age-specific earnings flows associated with multiple sources of income. We use this result in Section 3 of this paper.

As a by-product of Theorem 2, we can construct the distributions of $Y(t)$ for a variety of counterfactual histories leading up to t. Define a process based on the index crossing property for $I(t)$ without any requirement on the positivity or negativity of $I(t - j)$, $j > 0$. Let $B(t) = \mathbf{1}[I(t) \geq 0]$ where $B(t) \in \{0, 1\}$. Let $B^t = (B(1), \ldots, B(t))$ where b^t is defined as the vector of possible values of $B(t)$. $D(t)$ was defined as $B(t)$ given $D^t = (0)$. $B(t)$ is defined without this restriction.

With the $B(t)$ it is possible to contemplate many alternative histories ruled out in the

construction of $D(t)$. From Theorem 2, we can construct

$$\Pr\left(Y(t) \leq y(t) \mid B^t = b^t, X = x, Z = z\right)$$

for all of the 2^t possible sequences of B^t outcomes up to t including sequences that were ruled out in the definition of the model for $D(t)$ in equation (1) such as $b^t = (0, 1, 0, 1, \dots)$. We obtain these probabilities by reversing the $\Psi(t, Z)$ limits associated with the $\eta(1), \dots, \eta(t)$ arguments of equation (2).[28]

These counterfactuals are difficult to interpret if we take stopping time model (1) literally. They allow for the possibility of persons starting and stopping treatment on multiple occasions leading up to t. We can also identify the distribution of $Y(t)$ for persons who stop at some time after t $(T > t)$.[29] There are two ways to interpret these features of our model: (a) as a symptom of incomplete specification of the statistical model because it allows for reentry even though the economic model does not; or (b) as a desirable feature because it allows for richer specifications of the economic model that permit reentry.

Note further that the counterfactuals that are identified by fixing different $D(j)$ at different values have an asymmetric aspect. We can generate $Y(t)$ distributions for persons who are treated at t or before. Without further structure, we cannot generate the distributions of these random variables for people who receive treatment at times after t.

The source of this asymmetry is the generality of duration model (1). At each stopping time t, we acquire a new random variable $\eta(t)$ which can have arbitrary dependence with $Y(t)$ and $Y(t')$ for all t and t'. From Theorem 2, we can identify this dependence between $\eta(t)$ and $Y(t')$ if $t' \leq t$. We cannot identify the dependence between $\eta(t)$ and $Y(t')$ for $t' > t$ without imposing further structure on the unobservables.[30]. Thus we can identify the distribution of college outcomes for high school graduates who do not go on to college and

[28] Cunha, Heckman, and Navarro (2005d) develop a semiparametric ordered choice model with stochastic thresholds that rules out these extraneous sequences but at the price of eliminating option values from the dynamic discrete choice model.

[29] This is the event associated with $B^t = (0)$.

[30] One possible structure is a factor model which we apply to this problem in the next section.

can compare these to outcomes for high school graduates, so we can identify the parameter "treatment on the untreated." However, we cannot identify the distribution of high school outcomes for college graduates (*e.g.* treatment on the treated parameters) without imposing further structure.[31] Since we can identify the marginal distributions under the conditions of Theorem 2, we can identify pairwise average treatment effects for all t, t'.

Appendix C contrasts the model identified by Theorem 2 with a conventional static multinomial discrete choice model with an associated system of counterfactuals. In that Appendix, we prove semiparametric identification of the conventional static model of discrete choice joined with counterfactuals and show how to identify all of the standard counterfactuals. For that model there is a fixed set of unobservables governing all stopping times. Thus we do not acquire new unobservables associated with each stopping time. With suitable normalizations, we can identify the joint distributions of choices and associated outcomes without the difficulties, just noted, that appear in the reduced form dynamic model.

A Model for Discrete Outcome Counterfactuals

We next develop a discrete outcome analog to the results just presented for continuous outcomes. In this subsection, we suppose that associated with each stopping time at age a is a binary variable $e\left(a,t,X\right)$, denoting, for example, employment at age a for a person with stopping time (treatment time) $T=t$ with regressors X. For specificity, in the schooling example, treatment time t is the age at which a person drops out of school. We assume that $e\left(a,t,X\right)=\mathbf{1}\left[e^{*}\left(a,t,X\right)\geq 0\right], t=1,\ldots,\bar{T}, a=1,\ldots,\bar{A}$ where $e^{*}\left(a,t,X\right)=\mu^{e}\left(a,t,X\right)-U^{e}\left(a,t\right)$ and each $U^{e}(a,t)$ has zero mean and finite variance. In the schooling example we can think of the $e\left(a,t,X\right)$ as employment indicators before schooling is finished and after, for people who have exactly t years of schooling. In the following, we will make use of the condensed forms $e(t,X)$, $e\left(X\right)$, $e^{*}\left(t,X\right)$, $e^{*}\left(X\right)$, $\mu^{e}\left(t,X\right)$, $\mu^{e}\left(X\right)$, $U^{e}(t)$ and U^{e} as

[31] In the schooling example, we can identify treatment on the treated for the final category $\bar{T}$ since $D^{\bar{T}-1}=(0)$ implies $D\left(\bar{T}\right)=1$. Thus at stage $\bar{T}-1$, we can identify the distribution of $Y\left(\bar{T}-1\right)$ for persons for whom $D\left(0\right)=0,\ldots,D\left(\bar{T}-1\right)=0,D\left(\bar{T}\right)=1$. Hence if college is the terminal state and high school the state preceding college, we can identify the distribution of high school outcomes for college graduates.

described in Section 2.3. We assume $U^e(t) \perp\!\!\!\perp X, Z$. The $e(t,x)$ are factuals for $T=t$ and counterfactuals for stopping times other than t. Instead of analyzing only the outcome at t, we analyze the entire path of outcomes associated with stopping time t.

Ignoring the selection problem, identification of $\mu^e(X)$ (up to scale) is a standard application of known results in the semiparametric discrete choice literature. The scales are arbitrary because the inequality that generates $e(a,t,X)$ remains valid if the arguments are scaled by any positive constant. Let $\Psi^t(Z) = (\Psi(1,Z(1)),\ldots,\Psi(t,Z(t))$ and recall that $\eta^t = (\eta(1),\ldots,\eta(t))$. We prove the following theorem.

Theorem 3. *Assume data on $e(t,X), X, Z$ given $T=t$. Assume data on stopping times T and Z from a random sample across observations and that the T are not censored. Further assume that $\Psi^t(Z)$ and $\mu^e(t,X)$ are members of the Matzkin class of functions and that*

(i) *$(U^e(t),\eta^t)$ are continuous random variables with zero means, finite variances and with support $Supp(U^e(t))\times Supp(\eta^t)$ with upper and lower limits $\left(\bar{U}^e(t),\bar{\eta}^t\right)$ and $\left(\underline{U}^e(t),\underline{\eta}^t\right)$ respectively. These conditions hold for each subcomponent of each subvector. The joint system is thus measurably separable for each component with respect to every other component.*

(ii) *$(U^e(t),\eta^t) \perp\!\!\!\perp (X,Z)$, $t=1,\ldots,\bar{T}$,*

(iii) *$Supp(\mu^e(t,X),\Psi^t(Z)) = Supp(\mu^e(t,X)) \times \prod_{j=1}^{t} Supp(\Psi(j,Z(j)))$, $t=1,\ldots,\bar{T}$, and this holds for each component of each vector,*

(iv) *$Supp(\mu^e(t,X),\Psi^t(Z)) \supseteq Supp(U^e(t),\eta^t)$, $t=1,\ldots,\bar{T}$,*

(v) *$Supp(U^e(t),\eta^t) = Supp(U^e(t)) \times \prod_{j=1}^{t} Supp(\eta(j))$, $t=1,\ldots,\bar{T}$, and this holds for each component of each vector,*

Then we can identify $\Psi^t(Z)$, $\mu^e(t,X)$ and the joint distributions of $(U^e(t),\eta^t)$ under the Matzkin conditions applied to each component of $\mu^e(a,t,X)$, $U^e(a,t)$ and to each component

of $\Psi^t(Z)$ and the corresponding component of η^t. Applying the Matzkin conditions for the functions and random variables up to scale, we obtain the functions and the distributions of the random variables up to scale.

Proof. The proof for this case parallels that of Theorem 2 with two exceptions. Since we do not observe $e^*(t, X)$, but just its dichotomization, we cannot use its variation to trace out the distribution of $U^e(t)$, as we did with $y(t)$ in Theorem 2 to produce the desired variation with condition *(iv)* of Theorem 3. To substitute for this variation, we invoke condition *(iv)*. See Appendix D for the proof for the general case. We analyze the entire lifecycle path of the $e(a, t, X)$ instead of just the period t outcome. ■

In this setup, we can analyze strings of binary outcome sequences associated with each treatment time. Theorem 3 can be modified to cover the case of counterfactual durations and we sketch this extension in Corollary 2 below. Note that Theorem 3 is more general than Theorem 2 in the sense that we identify the model generating vector $e(t, X)$ and not just a scalar outcome like $Y(t)$. Theorem D.1 in Appendix D extends Theorems 2 and 3 to consider both cases and a vector version of $Y(t)$, as well as an associated measurement system.

To produce a result on semiparametric identification of a discrete time analogue of the Abbring and Van Den Berg (2003) model of counterfactuals for durations, we assemble ingredients from Theorems 1, 2 and 3. Let $\Delta(a, t, X)$ be an indicator of whether a person at age a, treatment time t and characteristics X is in a spell of the outcome being studied (*e.g.* of employment or unemployment). Individuals receive at most one treatment. Assume that $\Delta(0, t, X) = 0$ for all $t > 0$. A person starting in "0" exits to "1". We normalize the initial age to zero so the scales for measuring age and time of treatment are the same. The age where Δ first becomes 1 is the length of the initial spell and the treatment time is t.[32]

Let $\Delta^*(a, t, X) = \chi(a, t, X) - \nu(a, t)$ denote a latent variable where $\nu(a, t)$ has a zero mean and finite variance and $\nu(a, t) \perp\!\!\!\perp (X, Z)$ for all a, t. We use the condensed form

[32]Recall that exit events in period t occur instantaneously at the beginning of the period.

notation introduced in Section 2.3. In particular, we let $\nu(t) = \left(\nu(1,t), \ldots, \nu\left(\bar{A},t\right)\right)$, and $\chi(t,X) = \left(\chi(1,t,X), \ldots, \chi\left(\bar{A},t,X\right)\right)$. We define the indicator of remaining in the initial state at age a for treatment time t as

$$\Delta(a,t,X) = \mathbf{1}\left[\Delta^*(a,t,X) \geq 0\right] \quad \text{for} \quad \Delta^{a-1}(X) = (0)$$

where $\Delta^{a-1}(X)$ is the history of the process up through age $a-1$. To parallel the analysis of Abbring and Van Den Berg (2003), we consider flow sampling of new spells. Thus in an analysis of unemployment, individuals start unemployed and are unemployed at least through treatment, are treated at age a' (or time $t = a'$), and then are followed after treatment at least until they leave the initial state. Treatment time (or age) a' is the age in the spell at which training is administered.

Implicit in the treatment time decision rule is the requirement that an individual be in the starting state (0) in order to receive treatment. Thus for $T = t$, it is required that $\Delta(a,t,X) = 0$ for all $a \leq t$. Treatment is assumed to be instantaneous but under a nonanticipation assumption any effects of treatment are found in periods $a > t$. We can prove the following Corollary of Theorem 3.

Corollary 2. *Assume data on $\Delta(t,X), X, Z$ given $T = t$. Assume data on stopping times T and Z from an initial random sample of persons in the state "0". Further assume that $\Psi^t(Z)$ and $\chi(t,X)$ are members of the Matzkin class of functions and that*

(i) $(\nu(t), \eta^t)$ are continuous random variables with zero means, finite variances and support $Supp\left(\nu(t)\right) \times Supp\left(\eta^t\right)$ with upper and lower limits $(\bar{\nu}(t), \bar{\eta}^t)$ and $\left(\underline{\nu}(t), \underline{\eta}^t\right)$ respectively, for all $t = 1, \ldots, \bar{T}$. These conditions hold for each subcomponent of each subvector. The joint system is thus measurably separable for each component with respect to every other component.

(ii) $(\nu(t), \eta^t) \perp\!\!\!\perp (X, Z)$, for all $t = 1, \ldots, \bar{T}$.

(iii) $Supp\left(\chi\left(t,X\right),\Psi^{t}\left(Z\right)\right)=Supp\left(\chi\left(t,X\right)\right)\times\prod_{\ell=1}^{t}Supp\left(\Psi\left(\ell,Z(\ell)\right)\right)$, *for all* $t=1,\ldots,\bar{T}$, *and this holds for each component of each vector.*

(iv) $Supp\left(\chi\left(t,X\right),\Psi^{t}\left(Z\right)\right)\supseteq Supp\left(\nu(t),\eta^{t}\right)$, *for all* $t=1,\ldots,\bar{T}$.

(v) $Supp\left(\nu(t),\eta^{t}\right)=Supp\left(\nu\left(t\right)\right)\times\prod_{\ell=1}^{t}Supp\left(\eta\left(\ell\right)\right)$, *for all* $t=1,\ldots,\bar{T}$, *and this holds for each component.*

Then, under the Matzkin conditions, we can identify $\Psi^{t}\left(Z\right)$ *and* $\chi\left(t,X\right)$ *and the joint distributions of* $\left(\nu\left(t\right),\eta^{t}\right)$ *for* $t=1,\ldots,\bar{T}$. *If we weaken these conditions so that the class of the functions is only known up to scale, we identify these functions up to scale and distributions of the random variables up to scale.*

Proof. The proof uses the ingredients of Theorem 3 and for the sake of brevity is deleted. ■

The basic idea underlying the proof is that with sufficient variation in (X,Z), we can identify subsets of persons who survive in the initial state of unemployment untreated to any given age a with a high probability. Some of the previously untreated survivors are treated at a and followed at least until they leave "0". The model is intrinsically complex, requiring that the analyst correct for selection into various pre-treatment survivorship statuses. The analyst must also correct for the effect of survival up to a on the possibility of treatment at a. We do not develop the full model of treatment times for the reduced form duration analysis in this paper.[33]

Theorem 3 and Corollary 2 reverse the order of the *B-D* conditioning discussion presented in the previous subsection. Both exploit the properties of index models. The duration models for time to treatment or for time to exiting unemployment place restrictions on the order in which thresholds are permitted to cross zero and their dependence on survival times.

In the reduced form models for $Y\left(t\right),e\left(a,t\right)$ or $\Delta\left(a,t\right)$, the pre-treatment outcomes at each age can differ depending on the time of treatment even after controlling for the

[33]The model of treatment times in Abbring and Van Den Berg is also implicit.

X, the Z^t, the $U(t)$ and the η^t. Thus we do not have to impose the "no anticipations" assumption invoked by Abbring and Van Den Berg (2003) which requires that controlling for the variables in their model analogous to our X, Z^t, η^t and $U(t)$, the age a outcomes $(a < t)$ be the same for all treatment times after a. This requirement rules out that the future can cause the past and is an intuitive requirement of a causal model.[34] As we discuss in Section 3, this is an artifact of the incompleteness of the specification of reduced form models. This possibility arises because the framework in this section, like the framework of many reduced form models, is not sufficiently rich to model or identify the information sets of agents. Conditioning on the same information set, the outcomes at pretreatment age a $(a < t)$ are the same for persons with different treatment times as we show in the structural models of Section 3.[35]

The models for binary strings and durations also share the property with the model produced by Theorem 2 that counterfactuals for impossible strings of treatment time histories can be generated. This is a consequence of the index function structure. Recall our discussion of the B-D conditioning in the preceding subsection.

We now turn to the development of factor models that allow us to construct the joint distributions of outcomes across stopping times.

2.5 Using Factor Models to Identify Joint Distributions of Counterfactuals

From Theorem 2 and Theorem 3 or their generalization, Theorem D.1 in Appendix D, we can identify joint distributions of outcomes for each treatment time t and the index

[34] The requirement is imposed by requiring either that $Y(a,t) = Y(a,t')$ $[e(a,t) = e(a,t'); \Delta(a,t) = \Delta(a,t')]$ for all $\min(t',t) > a$, or the weaker requirement that the pretreatment distributions be the same. We note that in quantum electrodynamics, Feynman's equations explicitly predict that the future causes the past so a "common sense" notion of causality is violated in this branch of physics. See www.qedcorp.com/pcr/pcr/m13.html.

[35] In a perfectly certain environment, the "no anticipations" condition is meaningless since the treatment time is in the agent's information set and it is not possible to standardize information sets across people with different treatment times.

generating treatment times. We cannot identify the joint distributions of outcomes across treatment times. As a consequence, we cannot, in general, identify treatment on the treated parameters.[36]

Aakvik, Heckman, and Vytlacil (2005) and CHH show how to use factor models to identify the joint distributions across treatment times and recover the standard treatment parameters. We can use their approach to identify the joint distribution of $Y = (Y(1), \ldots, Y(\bar{T}))$.

The basic idea underlying this approach is to use distributions for outcomes measured at each treatment time t and on the choice index to construct the joint distribution of outcomes across treatment choices. To illustrate how the idea works, suppose that we augment Theorem 2 by appealing to Theorem D.1 in Appendix D to identify the joint distribution of the vector of outcomes at each stopping time along with $I^t = (I(1), \ldots, I(t))$ for each t. For each t, we may write

$$\begin{aligned} Y(a, t, X, U(a, t)) &= \mu(a, t, X) + U(a, t) \qquad a = 1, \ldots, \bar{A} \\ I(t) &= \Psi(t, Z) + \eta(t). \end{aligned}$$

From the Matzkin conditions, the scale is determined. If we specify the Matzkin functions only up to scale, we determine the functions up to scale and make a normalization. From Theorem 2 and Theorem D.1, we can identify the joint distribution of $(\eta(1), \ldots, \eta(t), U(1, t), \ldots, U(\bar{A}, t))$.

Suppose that we adopt a one factor model where θ is the factor. It has mean zero and we can represent the errors by

$$\begin{aligned} \eta(t) &= \varphi_t \theta + \varepsilon_{\eta(t)} \\ U(a, t) &= \alpha_{a,t} \theta + \varepsilon_{a,t}, \qquad a = 1, \ldots, \bar{A}, \qquad t = 1, \ldots, \bar{T}. \end{aligned}$$

The θ are independent of all of the $\varepsilon_{\eta(t)}$, $\varepsilon_{a,t}$ and the ε's are mutually independent mean

[36] In the schooling model, we can identify these parameters at terminal treatment time $\bar{T}$.

zero disturbances. The φ_t and $\alpha_{a,t}$ are called factor loadings. Since θ is an unobservable, its scale is unknown. We can set the scale of θ by normalizing one factor loading, say $\alpha_{\bar{A},\bar{T}} = 1$. From the joint distribution of $\left(\eta, U\left(\bar{T}\right)\right)$, we can form the covariances

$$\begin{aligned} Cov\left(U\left(a,\bar{T}\right),U\left(a',\bar{T}\right)\right) &= \alpha_{a,\bar{T}}\alpha_{a',\bar{T}}\sigma_\theta^2 \qquad a \neq a'. \\ Cov\left(U\left(a,\bar{T}\right),\eta(t)\right) &= \alpha_{a,\bar{T}}\varphi_t\sigma_\theta^2. \end{aligned}$$

For $\bar{A} \geq 3$, we can identify σ_θ^2, $\alpha_{a,t}, \varphi_t, a = 1, \ldots, \bar{A}$ for $t = 1, \ldots, \bar{T}$.[37] From this information we can form for $a \neq a'$ or $t \neq t''$ or both,

$$Cov\left(U\left(a,t\right),U\left(a',t''\right)\right) = \alpha_{a,t}\alpha_{a',t''}\sigma_\theta^2,$$

even though we do not observe outcomes for the same person at two different stopping times. Thus we can construct the joint distribution of $(U,\eta) = \left(U\left(1\right),\ldots,U\left(\bar{T}\right),\eta\right)$. From this joint distribution we can recover the standard mean treatment effects as well as the joint distributions of the potential outcomes. We can determine the percentage of participants at treatment time t who benefit from participation compared to what their outcomes would be at other treatment times. We can perform a parallel analysis for the index functions $e^*(a,t,X)$ used to generate $e(a,t,X)$ in Section 2.4 as well as for the $\Delta^*\left(a,t,X\right)$. Conventional factor analysis assumes that the unobservables are normally distributed. CHH establish nonparametric identifiability of the θ's and the ε's and their analysis of nonparametric identifiability applies here.

[37] **Proof.** Assume that the factor loadings and variances are nonzero. From the normalization it follows that

$$\frac{Cov\left(U\left(a,\bar{T}\right),U\left(a',\bar{T}\right)\right)}{Cov\left(U\left(a,\bar{T}\right),U\left(\bar{A},\bar{T}\right)\right)} = \alpha_{a',\bar{T}},\ a' = 1,2,\ldots,\bar{A};\ \bar{A} \geq 3.$$

$$Cov\left(U\left(\bar{A},\bar{T}\right),U\left(a',\bar{T}\right)\right) = \alpha_{a',\bar{T}}\sigma_\theta^2$$

Since we know $\alpha_{a',\bar{T}}$, we can identify σ_θ^2. We can identify $\varphi_t, t = 1,\ldots,\bar{T}$ from

$$Cov\left(U\left(a,t\right),\eta(t)\right) = \alpha_{a,t}\varphi_t\sigma_\theta^2.$$

■

In the schooling example discussed in the previous subsection, having access to these distributions means that we can form not only the potential earnings in college of a high school graduate as we could without invoking the factor structure assumption, but we are also able to generate the distribution of potential earnings in high school of a college graduate. Thus, in addition to the pairwise average treatment effects that can be formed using the output of Theorem 2, we can form treatment on the treated, as well as all of the distributional treatment effects discussed in CHH, Heckman and Smith (1998) and Heckman and Vytlacil (2006a). As noted by CHH and Cunha, Heckman, and Navarro (2005a,b,c,e), we can also form the joint distribution of college and high school earnings for college graduates.

Theorem 2, strictly applied, actually produces only one scalar outcome for each stopping time. We need three or more measurements for each stopping time to use factor analysis. Theorem D.1 in Appendix D extends the analysis of Theorem 2 to a vector outcome case. If vector outcomes are not available, access to a measurement system M that assumes the same values for each stopping time can substitute for the need for vector outcomes for Y. Let M_j be the j^{th} component of this measurement system. Write

$$M_j = \mu_{j,M}(X) + U_{j,M}, \quad j = 1, \ldots, J,$$

where $U_{j,M}$ are mean zero and independent of X.

Suppose that the $U_{j,M}$ have a one-factor structure so $U_{j,M} = \alpha_{j,M}\theta + \varepsilon_{j,M}$, $j = 1, \ldots, J$, where the $\varepsilon_{j,M}$ are mean zero, mutually independent random variables, independent of the θ. Adjoining these measurements to the one outcome measure $Y(t)$ with a factor structure joined with two or more measurements ($J \geq 2$) can substitute for the measurements of $Y(a,t)$ used in the previous example. In an analysis of schooling, the M_j can be test scores that depend on ability θ. Ability is assumed to affect outcomes $Y(t)$ and the choice of treatment times indices arrayed in I.

These examples illustrate the wealth of counterfactual within—and across—stopping time

t distributions that can be produced from the factor models developed in Aakvik, Heckman, and Vytlacil (2005) and in CHH. The factor models are convenient vehicles for generating low-dimensional representations of unobservables. Alternative methods for generating low dimensional representations of unobservables that can be used to construct counterfactual distributions across treatment times are pursued in Urzua (2005).

Factor models generalize the method of matching. Conditional on θ,X, Z, all of the potential outcomes are independent of $D(l)$: $Y(t) \perp\!\!\!\perp D(l) \mid X, Z, \theta$ for all $t, l = 1, \ldots, \bar{T}$. Our analysis allows for the possibility that θ is unobserved by the economist. The price of allowing for this is assuming that $\theta \perp\!\!\!\perp (X, Z)$. This assumption is not required in matching, if we observe the θ.[38]

A limitation of the reduced form approach pursued in this section is that, because the underlying model of choice is not clearly specified, it is not possible without further structure to form, or even define, the marginal treatment effect analyzed in Heckman and Vytlacil (1999, 2001, 2005, 2006a,b) or Heckman, Urzua, and Vytlacil (2005). The absence of well defined choice equations is problematic for the model we have analyzed thus far, although it is typical of many statistical treatment effect analyses.[39,40] In this framework, it is not possible to distinguish objective outcomes from subjective evaluations of outcomes, and to distinguish *ex ante* from *ex post* outcomes. It is also possible to identify counterfactuals that can depend on future treatment times, contrary to intuitions that the future cannot cause the past. We can rule out such models by assumption as is the practice in the statistical treatment effect literature (see *e.g.* Robins, 1997; Gill and Robins, 2001; Lok, 2001) but the assumptions on the underlying economic model required to do this are not clearly articulated

[38] Conditioning on observables to produce conditional independence models between counterfactuals and assignment is discussed in Rosenbaum and Rubin (1983), Gill and Robins (2001), Lechner and Miquel (2002), Heckman and Navarro (2004), and CHH.

[39] Heckman (2006) and Heckman and Vytlacil (2006a,b) point out that one distinctive feature of the economic approach to program evaluation is the use of choice theory to define parameters and evaluate alternative estimators.

[40] This contrasts with the semiparametric model for treatment effects in a multinomial choice model in Appendix B, where a well defined choice criterion exists. This appendix defines $EOTM$, the effect of treatment for people at the margin, for a classical multinomial choice model with associated outcomes. See also CHH.

in the reduced form approach.

We now develop an explicit economic model for dynamic treatment effects that allows us to make these and other distinctions and to eliminate hard-to-interpret features of the statistical model. We extend the analysis presented in this section to a more precisely formulated economic model. We explicitly allow for agent updating of information sets. A well posed economic model rules out the possibility that the future causes the past as part of the model specification. It also enables us to evaluate policies in one environment and accurately project them to new environments as well as accurately forecast new policies never previously experienced. See Heckman (2006) and Heckman and Vytlacil (2005, 2006a,b).

3 A Sequential Structural Model with Option Values

This section analyzes the identifiability of a structural sequential optimal stopping time model. We use ingredients assembled in the previous sections to build an economically interpretable framework for analyzing dynamic treatment effects. We focus on a schooling model with associated earnings outcomes that is motivated by the work of Keane and Wolpin (1997) and Eckstein and Wolpin (1999). We explicitly model costs and allow for work while in school. We allow for the arrival of serially correlated shocks in information more general than those entertained by Pakes (1986), Rust (1987), Hotz and Miller (1993), Manski (1993), Keane and Wolpin (1997) or Eckstein and Wolpin (1999).

In the model of this section it is possible to interpret the literature on dynamic treatment effects within the context of an economic model; to allow for earnings while in school as well as grade-specific tuition costs; to separately identify returns and costs; to distinguish private evaluations from "objective" *ex ante* and *ex post* outcomes and to identify persons at various margins of choice. In the context of medical economics, we consider how to identify the pain and suffering associated with a treatment as well as the distribution of benefits from the intervention. We also model how anticipations about potential future outcomes associated

with various choices evolve over the life cycle as sequential treatment choices are made.

In contrast to the analysis of Section 2, the identification proof for our dynamic choice model works in reverse starting from the last period and sequentially proceeding backward. This approach is required by the forward-looking nature of dynamic choice analysis and makes an interesting contrast with our reduced form analyses which proceed forward from initial period values.

We use limit set arguments to identify the parameters of outcome and measurement systems for each stopping time $t = 1, \ldots, \bar{T}$, including means and joint distributions of unobservables. These systems are identified without invoking any special assumptions about the structure of model unobservables. If we invoke factor structure assumptions for the unobservables, we identify the factor loadings associated with the measurements (as defined in Section 2.5) and outcomes. We also nonparametrically identify the distributions of the factors and the distributions of the innovations to the factors. With the joint distributions of outcomes and measurements in hand for each treatment time, we can identify cost (and preference) information from choice equations that depend on outcomes and costs (preferences). We can also identify joint distributions of outcomes across stopping times. Thus we can identify the proportion of people who benefit from treatment. Our analysis generalizes the one shot decision models of Cunha, Heckman, and Navarro (2005a,b,c,e) to a sequential setting.

Because our model makes many new distinctions that are not possible in the analysis of Section 2, we have to introduce some new notation. Agents make decisions about schooling at each age in their life cycle, and we are explicit about their decision rule.

Agents sequentially select schooling levels. New information arrives at each stage. One of the benefits of continuing on in a process is the arrival of new information. Let $t(a) \in \{1, \ldots, \bar{T}\}$ index the schooling level that an individual has attained at age $a \in \{1, \ldots, \bar{A}\}$. The person may go on to attain more years of schooling. Each year of schooling takes one year of age to complete. We assume that there is no grade repetition and we assume that

once persons stop schooling, they never return. It would be better to derive such stopping behavior as a feature of a more general model with possible recurrence of states but we do not do so here.[41]

As a consequence of these assumptions, $t(a) = a$ up to the time the person drops out of school. Aging continues but schooling does not. We set $D(a) = 0$ if the individual decides to continue to the next level of schooling (*i.e.*, does not receive "treatment" at age a) and $D(a) = 1$ if the individual stops at a. In our notation, final schooling level (time at treatment) $T = t(a)$ is the first age a (grade $t(a)$) at which $D(a) = 1$. Equivalently, we could denote this event by $D(t(a)) = 1$, because up to the time of dropout from the schooling process $a = t(a)$. Individuals start life in the schooling state $D(0) = 0$. Define $\delta(a) = 1 - \mathbf{1}\left[\sum_{j=0}^{a-1} D(j) = 0\right]$ to be an indicator of whether the individual stopped (received treatment) by age a (so $\delta(a) = 1$) or whether the individual is still a student entering age a (so $\delta(a) = 0$).[42] Figure 1 shows the evolution of age and grades, and clarifies the notation.

Let individual earnings at age a for a person with current schooling level $t(a)$ be written as

$$Y(a, t(a), \delta(a), X) = \mu(a, t(a), \delta(a), X) + U(a, t(a), \delta(a)), \tag{3}$$

so $Y(a, t(a), 0, X)$ denotes the earnings of an individual with characteristics X who is still enrolled in school and goes on to complete at least $t(a)+1$ years of schooling. $U(a, t(a), \delta(a))$ is a mean zero shock that is unobserved by the econometrician but may, or may not, be observed by the agent. It is the earnings of the person as a student at age a. $Y(a, t(a), 1, X)$ denotes the earnings at age a of an agent who has decided to stop schooling at or before age a. When $\delta(a) = 1$, $Y(a, t(a), \delta(a), X)$ is meaningfully defined only if $a \geq t(a)$. We impose this restriction throughout, and define all counterfactuals invoking this assumption to produce interpretable models.

[41] See Heckman, Urzua, and Yates (2005) for the derivation identification and estimation of such a model.

[42] Recall that treatment is instantaneous and occurs at the start of the period.

The direct cost of remaining enrolled in school at age a is

$$C\left(t(a), X, Z\left(t(a)\right)\right) = \Phi\left(t(a), X, Z\left(t(a)\right)\right) + W\left(t(a)\right)$$

where X and $Z\left(t\left(a\right)\right)$ are vectors of observed characteristics (from the point of view of the econometrician) that affect costs at schooling level $t(a)$, and $W(t(a))$ are mean zero shocks that are unobserved by the econometrician that may or may not be observed by the agent. Costs are paid in the period before schooling is undertaken. The agent is assumed to know the costs of making schooling decisions at each transition. The agent is also assumed to know the X and the $Z(t(a))$ for all periods.[43]

Once an agent decides to stop at schooling level $T = t$, she never returns to school. Under these assumptions, the expected reward at age a of stopping (*i.e.*, receiving treatment) at $T = t$ is given by the expected present value of her remaining lifetime earnings:

$$R\left(a, t, X\right) = E\left(\sum_{j=0}^{\bar{A}-t}\left(\frac{1}{1+r}\right)^{j} Y\left(a+j, t, 1, X\right) \;\middle|\; \mathcal{I}_a\right), \tag{4}$$

where $\mathcal{I}_a$ is the age-specific information set which includes the schooling level attained at age a as well as all state variables known to the agent including conditional distributions of future variables that are forecast by the agent. A more accurate notation would write $R\left(a, t, \mathcal{I}_a\right)$ but it is convenient in the proofs to use $R\left(a, t, X\right)$ and we do so in this paper. We assume a fixed, nonstochastic, interest rate r.[44] Because agents are forward looking, we define the cost shifters for schooling levels $t\left(a\right)$ and beyond as $Z^{t(a)} = \left(Z\left(t\left(a\right)\right), Z\left(t\left(a\right)+1\right), \ldots, Z\left(\bar{T}-1\right)\right)$, and define the entire vector of cost shifters as $Z = Z^1$. Agents are assumed to know these cost shifters and they are in the information set $\mathcal{I}_a$. The continuation value at age a and schooling level $t\left(a\right)$ given X and $Z^{t(a)}$ is denoted $K\left(a, t\left(a\right), \mathcal{I}_a\right)$.

At $\bar{T}-1$, when an individual decides whether to stop or continue on to $\bar{T}$, the expected

[43] These assumptions can be relaxed and are made for convenience. See Cunha, Heckman, and Navarro (2005e) for a discussion of selecting variables in the agent's information set.

[44] This assumption can be relaxed but we do not do so in this paper.

reward from remaining enrolled and continuing to $\bar{T}$ (*i.e.*, the continuation value) is the earnings while in school less costs plus the expected discounted future return that arises from completing $\bar{T}$ years of schooling:

$$\begin{aligned} K\left(\bar{T}-1,\bar{T}-1,\mathcal{I}_{\bar{T}-1}\right) &= Y\left(\bar{T}-1,\bar{T}-1,0,X\right)-C\left(\bar{T}-1,X,Z\left(\bar{T}-1\right)\right) \\ &\quad +\frac{1}{1+r}E\left(R\left(\bar{T},\bar{T},X\right)\mid\mathcal{I}_{\bar{T}-1}\right) \end{aligned}$$

where $C\left(\bar{T}-1,X,Z\left(\bar{T}-1\right)\right)$ is the direct cost of schooling for the transition to $\bar{T}$. This expression embodies our assumption that each year of school takes one year of age. $\mathcal{I}_{\bar{T}-1}$incorporates all of the information known to the agent.

The value function at $\bar{T}-1$ is the larger of the two expected rewards that arise from stopping at $\bar{T}-1$ or continuing one more period to $\bar{T}$:

$$V\left(a,\bar{T}-1,\mathcal{I}_{\bar{T}-1}\right)=\max\left\{R\left(\bar{T}-1,\bar{T}-1,X\right),K\left(\bar{T}-1,\bar{T}-1,\mathcal{I}_{\bar{T}-1}\right)\right\}.$$

More generally, at age a and schooling level $t\left(a\right)$, the value function is

$$\begin{aligned} V\left(a,t\left(a\right),\mathcal{I}_{t(a)}\right) &= \max\left\{R\left(a,t\left(a\right),X\right),K\left(a,t\left(a\right),\mathcal{I}_{t(a)}\right)\right\} \\ &= \max\left\{R\left(a,t\left(a\right),X\right),\left(\begin{array}{c} Y\left(a,t\left(a\right),0,X\right)-C\left(t\left(a\right),X,Z\left(t\left(a\right)\right)\right) \\ +\frac{1}{1+r}E\left(V\left(a+1,t\left(a\right)+1,\mathcal{I}_{t(a)+1}\right)\mid\mathcal{I}_{t(a)}\right) \end{array}\right)\right\}. \end{aligned}$$

The option value at age a of continuing schooling further than $t\left(a\right)$ is given by the difference between the reward an individual expects to obtain by going to one more year of school, taking into consideration that he might go even further, and the reward he expects to obtain

by stopping the next year,

$$\begin{aligned} O\left(a,t\left(a\right)\right) &= Y\left(a,t\left(a\right),0,X\right)-C\left(t\left(a\right),X,Z\left(t\left(a\right)\right)\right) \\ &\quad +\frac{1}{1+r}E\left(V\left(a+1,t\left(a\right)+1,\mathcal{I}_{t(a)+1}\right)\mid\mathcal{I}_{t(a)}\right) \\ &\quad -\left\{\begin{array}{l} Y\left(a,t\left(a\right),0,X\right)-C\left(t\left(a\right),X,Z\left(t\left(a\right)\right)\right) \\ +\frac{1}{1+r}E\left(R\left(a+1,t\left(a\right)+1,X\right)\mid\mathcal{I}_{t(a)}\right) \end{array}\right\}. \end{aligned}$$

There is no option value for persons who have completed schooling. Collecting terms, for $a=t\left(a\right)$,

$$\begin{aligned} O\left(a,t\left(a\right)\right) &= \frac{1}{1+r}E\left(V\left(a+1,t\left(a\right)+1,\mathcal{I}_{t(a)+1}\right)-R\left(a+1,t\left(a\right)+1,X\right)\mid\mathcal{I}_{t(a)}\right) \\ &= \frac{1}{1+r}E\left(\max\left\{R\left(a+1,t\left(a\right)+1,X\right),K\left(a+1,t\left(a\right)+1,\mathcal{I}_{t(a)+1}\right)\right\}\mid\mathcal{I}_{t(a)}\right) \\ &\quad -\frac{1}{1+r}E\left(R\left(a+1,t\left(a\right)+1,X\right)\mid\mathcal{I}_{t(a)}\right).^{45} \end{aligned}$$

In the notation for index functions introduced in Section 2, we define the decision rule using $I(a,t\left(a\right),\mathcal{I}_{t(a)})=R(a,t\left(a\right),X)-K(a,t\left(a\right),\mathcal{I}_{t(a)})$ where

$$D\left(a\right)=\mathbf{1}\left[I\left(a,t\left(a\right),\mathcal{I}_{t(a)}\right)>0,\ I(a-1,t(a)-1,\mathcal{I}_{t(a)-1})\leq 0,\ldots,I(1,1,\mathcal{I}_{1})\leq 0\right].$$

For proving identification, it is useful to separate out the component of the cost function based on observables (from the point of view of the econometrician), $\Phi(t\left(a\right),X,Z(t\left(a\right)))$,

[45] Our model allows no recall and is clearly a simplification of a more general model of schooling with option values. Instead of imposing the requirement that once a student drops out the student never returns, it would be useful to derive this property as a feature of the economic environment and the characteristics of individuals. In a more general model, different persons could drop out and return to school at different times as information sets are revised. This would create further option value beyond the option value developed in the text that arises from the possibility that persons who attain a given schooling level can attend the next schooling level in any future period. Implicit in our analysis of option values is the additional assumption that persons must work at the highest level of education for which they are trained. An alternative model allows individuals to work each period at the highest wage across all levels of schooling that they have attained. Such a model may be too extreme because it ignores the costs of switching jobs, especially at the higher educational levels where there may be a lot of job-specific human capital for each schooling level. A model with these additional features is presented in Heckman, Urzua, and Yates (2005).

from the rest of the index which include the unobservable $W(t(a))$ as well as other ingredients. We define a subindex of $I(a,t(a),\mathcal{I}_{t(a)})$ as following:

$$\begin{aligned}\Upsilon\left(t(a),X,Z^{t(a)+1}\right) &= R(a,t(a),X)-[Y(a,t(a),0,X)-W(t(a))\\ &\quad+\frac{1}{1+r}E\left(V\left(a+1,t(a)+1,\mathcal{I}_{t(a)+1}\right)\mid\mathcal{I}_{t(a)}\right)].\end{aligned}$$

Thus,

$$I\left(a,t(a),\mathcal{I}_{t(a)}\right)=\Phi\left(t(a),X,Z(t(a))\right)+\Upsilon\left(t(a),X,Z^{t(a)+1}\right),$$

where $\Upsilon(t(a),X,Z^{t(a)+1})$ is the "error term" of the index function generating the model. We use the notation $\Upsilon\left(t(a),X,Z^{t(a)+1}\right)$ because it is helpful to understand the argument of the proofs presented in the next section. However, a more accurate notation would be $\Upsilon\left(t(a),\mathcal{I}_{t(a)}\right)$ where $\mathcal{I}_{t(a)}$ is the information set of the agent at stage $t(a)$ which may include $Z^{t(a)}$ and X.

An individual stops at the schooling level at the first age where this index becomes positive.[46] From data on stopping times, we can nonparametrically identify the conditional probability of stopping at a,

$$\Pr(T=t(a)\mid X,Z)=\Pr\left(\begin{array}{l|l} I\left(a,t(a),\mathcal{I}_{t(a)}\right)>0, & \\ I\left(a-1,t(a)-1,\mathcal{I}_{t(a)-1}\right)\leq 0,\ldots, & X=x,Z=z\\ I(1,t(1),\mathcal{I}_1)\leq 0 & \end{array}\right),$$

where $a=t(a)$ until the age where a person stops schooling, and $\delta(a)=1$.

In order to identify the sequential revelation of information and to identify the cost functions, we represent the unobservables (from the point of view of the econometrician)

[46] This makes implicit assumptions about the economic environment facing agents. Stationarity of the environment would produce this outcome but it is only a sufficient condition. We leave development of more precise conditions for later work.

using a factor structure tailored to the notation of this section,

$$\left.\begin{array}{l} U\left(a,t\left(a\right),\delta\left(a\right)\right)=\theta\alpha_{a,t(a),\delta(a)}+\varepsilon(a,t(a),\delta\left(a\right)) \\ W\left(t\left(a\right)\right)=\theta\lambda_{t(a)}+\xi\left(t\left(a\right)\right) \end{array}\right\} \qquad a=1,\ldots,\bar{A},\ t(a)\leq a \text{ for } \delta(a)=1,$$

where the subscript on the factor loading is the argument of the variable being given a factor representation. We assume that the measurement equations (the M of Section 2.5) can also be factor analyzed using θ, an L-dimensional vector of factors $(\theta_1,\ldots,\theta_L)$ and that $U_{j,M}=\theta\alpha_{j,M}+\varepsilon_{j,M},\ j=1,\ldots,J$. We also assume that the ε and ξ have zero means and finite variances and are component-wise independent and independent of the θ which are also component-wise independent: $\theta_i \perp\!\!\!\perp \theta_j$, for all $i\neq j,\ i,j=1,\ldots,L$.[47]

The agent is assumed to make choices using rational expectations. By this we mean that the agent whose choice behavior is being analyzed knows the distributions of θ, $\{\varepsilon(a,\ t(a),\ \delta\left(a\right))\}_{a=1}^{\bar{A}}$, $\xi\left(t\left(a\right)\right)$ for all a and $t\left(a\right)=a,\ldots,\bar{T}-1$, and $\{\varepsilon_{j,M}\}_{j=1}^{J}$ and uses them in making choices. We assume that the parameters of the model as well as $X,\ Z,\ \{\xi\left(t\left(a\right)\right)\}_{t(a)=1}^{\bar{T}-1}$, $\{\varepsilon_{j,M}\}_{j=1}^{J}$ are known by the agent and are in the information set $\mathcal{I}_a$ but the values of $\varepsilon(a+k,t(a+k),\delta\left(a+k\right))$, $k>0$ are not. Agents may or may not know θ.

One possible specification of the information structure of the model regarding θ is the following.

(I-1) *At age a, each element of θ is either known to the agent or it is not known. Thus, when revelation about θ occurs, it is instantaneous.*

This assumption rules out gradual learning, such as standard Bayesian updating. We further assume that

[47] Thus we assume that

$$\theta_j \perp\!\!\!\perp \varepsilon(a,t(a),\delta\left(a\right)) \text{ for all } j,a,t\left(a\right),\delta\left(a\right);\ \varepsilon(a,t(a),\delta\left(a\right)) \perp\!\!\!\perp \varepsilon(a',t''(a'),\delta'''\left(a'\right)),\ \text{for all } a',t''(a'),\delta'''\left(a'\right);$$

$$\text{except if } a=a'',\ t''(a')=t\left(a\right) \text{ and } \delta\left(a\right)=\delta'''\left(a'\right);\ \varepsilon(a,t(a),\delta\left(a\right)) \perp\!\!\!\perp \xi\left(t\left(a\right)\right) \text{ for all } a,t\left(a\right),\delta\left(a\right);$$

$$\theta_j \perp\!\!\!\perp \xi\left(t\left(a\right)\right) \text{ for all } j,a;\ \theta_\ell \perp\!\!\!\perp \varepsilon_{j,M} \text{ for all } \ell=1,\ldots,L;\ j=1,\ldots,J.$$

(I-2) *Information arrives about the elements of θ sequentially (*e.g.*, in the first a_1 periods of earnings only the first element of θ enters, in the next a_2 periods the first two elements of θ enter, and so on). If the l^{th} element of θ affects earnings at $a_\ell \leq a$, then it is known by the agent at a_ℓ and ever after.*

These assumptions allow for the possibility that agents may know some or all the elements of θ at a given age a regardless of whether or not they determine earnings at or before age a. Once known, they are not forgotten. As agents accumulate information, they revise their forecasts of their future earnings prospects at subsequent stages of the decision process. This affects their decision rules and subsequent choices. Thus we allow for learning which can affect both pretreatment outcomes and posttreatment outcomes.[48] We use this specification in the empirical work reported in Heckman (2006). Other specifications of the updating of the information sets of agents are possible.[49] All dynamic discrete choice models make some assumptions about the updating of information and any rigorous identification analysis must test among competing specifications of information updating.

Variables unknown to the agent are integrated out by the agent in forming value functions. Variables known to the agent are treated as constants by the agents. They are integrated out by the econometrician. In general, the econometrician knows less than what the agent knows. The econometrician seeks to identify the distributions in the agent information sets that are used by the agents to form their expectations as well as the distributions of variables known to the agent and treated as certain quantities by the agent but not known by the econometrician. Determining which elements belong in the agent's information set can be done using the methods exposited in Cunha, Heckman, and Navarro (2005e) and Navarro (2004b) who consider testing what components of X, Z, ξ, ε as well as θ are in the agent's

[48] This type of learning about unobservables is ruled out in the Abbring – Van den Berg model (2003). However, in our model, conditioning on the same information set $\mathcal{I}_a$, the distributions of pretreatment costs and earnings are the same for all persons irrespective of their treatment times.

[49] It is fruitful to distinguish models with exogenous arrival of information (so that information arrives at each age a independent of any actions taken by the agent) from information that arrives as a result of choices by the agent. Our model is in the first class. The model of Miller (1984) or Pakes (1986) are in the second class.

information set. We briefly discuss this issue at the end of the next section.[50] We now establish semiparametric identification of the model assuming a given information structure. Determining the appropriate information structure facing the agent and its evolution is an essential aspect of identifying any dynamic discrete choice model.

Observe that agents with the same information sets at age a, $\mathcal{I}_a$, have the same expectations of future returns, and the same value functions. Persons with the same *ex ante* reward, state and preference variables have the same *ex ante* distributions of stopping times. *Ex post*, stopping times may differ among agents with identical *ex ante* information sets. Controlling for $\mathcal{I}_a$, future realizations of stopping times do not affect past rewards.

3.1 Semiparametric identification of dynamic sequential discrete choice models

Establishing semiparametric identifiability of our model is a nontrivial task because of its intrinsic nonlinearity. Our strategy is as follows. Using limit set arguments which we specify below, we can identify the joint distributions of earnings (for each treatment time t across a) and any associated measurements that do not depend on the stopping time chosen. For each stopping time, we can construct the means of earnings outcomes at each age and of the measurements and the joint distributions of the unobservables for earnings and measurements. Factor analyzing the joint distributions of the unobservables, under conditions specified in CHH and Navarro (2004a), we identify the factor loadings, and nonparametrically identify the distributions of the factors and the independent components of the error terms in the earnings and measurement equations. Armed with this knowledge, we can use choice data to identify the distribution of the components of the cost functions that are not directly observed. We can also construct the joint distributions of outcomes across stopping times.

To simplify the notation in our proofs, we use the condensed forms for the variables

[50] Our model of learning is clearly very barebones. Information arrives exogenously across ages. In the factor model, all agents who advance to a stage get information about additional factors at that stage of their life cycles but the realizations of the factors may differ across persons. Urzua (2005) extends this analysis.

$U(t,0), U(t,1), U(0), U(1), U, Y(t,0,X), Y(t,1,X), Y(0,X), Y(1,X), Y(X), \mu(t,0,X)$, $\mu(t,1,X)$, $\mu(0,X)$, $\mu(1,X)$, $\mu(X)$, W, $\Phi(X,Z)$ and $\Upsilon(X,Z)$ that were introduced in Section 2.3. We also define $M = (M_1, \ldots, M_J)$, $U_M = (U_{1,M}, \ldots, U_{J,M})$ and $\mu_M(X) = (\mu_{1,M}(X), \ldots, \mu_{J,M}(X))$. The X may have subvectors depending on time but to simplify the analysis we suppress the individual elements. We embed the restriction that when $\delta(a) = 1$, $a \geq t(a)$ and restrict ourselves to counterfactuals and factuals with arguments that satisfy this property.

Using this notation, we can state our identification strategy more precisely. Using limit sets that make the probability of each stopping time, $t = 1, \ldots, \bar{T}$, arbitrarily close to 1, we construct the joint distribution of $(Y(t,0,X), Y(t,1,X), M)$ including the joint distribution of $U(0,t)$, $U(1,t)$ and U_M. Using factor analysis, we determine the factor loadings, and identify the joint distribution of θ, ε, ξ nonparametrically. With the factor loadings and these distributions in hand, we can use choice data to identify the mean of the cost function in the terminal schooling choice ($\Phi(\bar{T}-1, X, Z)$) and the distribution of unobservable components of costs $W(\bar{T}-1)$. Backward inducting, we can identify the $\Phi(t, X, Z)$ and the distribution of $W(t)$ for the remaining transitions. Using our factor structure, we can identify the full joint distributions of *ex post* outcomes and measurements $(Y(1,X), Y(0,X), M)$ across stopping times. We first establish identification of the joint distribution of $(Y(t,0,X), Y(t,1,X), M)$ for each t.

Theorem 4. *Assume that*

(i) U, U_M *and* W *are continuous random variables with mean zero, finite variance and support* $Supp(U) \times Supp(U_M) \times Supp(W)$ *with upper and lower limits* $\bar{U}, \bar{U}_M, \bar{W}$ *and* $\underline{U}, \underline{U}_M, \underline{W}$*, respectively, which may be bounded or infinite. We assume that this condition applies to each component of* U*,* U_M*, and* W*, and all possible combinations of components. The cumulative distribution function of* $W(t(a)), t(a) = 1, \ldots, \bar{T}$ *is assumed to be strictly increasing over its full support* $(\underline{W}(t(a)), \bar{W}(t(a)))$*, for all* $t(a) = 1, \ldots, \bar{T}$*.*

(ii) $(X, Z) \perp\!\!\!\perp (U, U_M, W)$.

(iii) $Supp\left(\mu\left(X\right), \mu_M(X), \Phi\left(X, Z\right)\right) = Supp\left(\mu\left(X\right)\right) \times Supp(\mu_M(X)) \times Supp\left(\Phi\left(X, Z\right)\right)$ *and this holds element by element.*

(iv) $Supp\left(\Phi\left(X, Z\right)\right) \supseteq Supp\left(-\Upsilon\left(X, Z\right)\right)$ *and this holds element by element within each vector.*

Then, $\mu\left(a, t\left(a\right), \delta\left(a\right), X\right), \mu_M(X)$*, the joint distribution of* $\left(U\left(a, t\left(a\right), \delta\left(a\right)\right), U_M\right)$ *are identified.*[51]

Proof. Assumptions *(iii)* and *(iv)* are sufficient conditions for limit sets $\mathcal{Z}^1$ and $\mathcal{Z}^0$ to exist such that $\lim_{X,Z \to \mathcal{Z}^1} P(T = t\left(a\right) \mid X, Z) = 1$, *i.e.*, there is a limit set in which the individuals are observed to stop at a, $T = t\left(a\right)$ and hence $\delta\left(a\right) = 1$ with probability one, and $\lim_{X,Z \to \mathcal{Z}^0} P\left(T = t\left(a\right) \mid X, Z\right) = 0$, so that there is a limit set of individuals who remain in school at $a = t\left(a\right)$ so $\delta\left(a\right) = 0$ with probability one in that limit set. One way to satisfy this condition is through exclusion restrictions: having an element in each $Z(j)$, call it $Z^*(j)$, that is not in X or $Z(j')$, $j \neq j'$, assuming that the $\Phi\left(j, X, Z(j)\right)$ is monotonic in $Z^*(j)$, and assuming that the $Z^*(j)$ can be independently varied, conditional on all of the remaining Z and the X. Since future costs enter this probability, we can potentially use any argument of $\Phi\left(t\left(a'\right), X, Z\left(t\left(a'\right)\right)\right)$, $a' > a$ to obtain these limits. Furthermore, with time varying components of X, some elements of future X might be available to achieve the required variation, provided support conditions are met. Under the limit set assumption, identification of $\mu\left(a, t\left(a\right), \delta\left(a\right), X\right)$, $\mu_M(X)$ and the marginal distribution of $\left(U\left(a, t\left(a\right), \delta\left(a\right)\right), U_M\right)$ follows immediately.

Identification of the joint distribution of (U_M, U) follows from the fact that, in the limit set, we can form the left hand side of

$$\Pr\left(U_M < m - \mu_M\left(X\right), U < y - \mu\left(X\right) \mid X = x\right) = F_{U_M,U}\left(m - \mu_M\left(x\right), y - \mu\left(x\right)\right).$$

[51] Recall that we restrict the admissible counterfactuals to have arguments that satisfy $a > t(a)$ when $\delta(a) = 1$.

We can trace out this distribution by finding vectors q_1 and q_2 defined so $q_1 = m - \mu_M(X)$ and $q_2 = y - \mu(X)$ and by independently varying the points of evaluation of the components of q_1 and q_2. ■

This theorem applies to any known transformation of the Y and M that satisfies the property of separability of the errors. Notice that we do not need conventional exclusion restrictions to identify the objects produced from Theorem 4. Notice further that we do not need to invoke the Matzkin conditions or the linearity-in-parameters conditions for the cost function to secure identification of the joint distribution of outcomes and measurements for each stopping time.

From this theorem, we can produce average treatment effects for outcomes for any pair of stopping times.[52] To produce the joint distribution of outcomes across stopping times, we can use factor analysis applied to the joint distribution as described in Section 2.5. Under conditions on the unobservables specified in CHH and Navarro (2004a), we can nonparametrically identify the distribution of the factors and the uniquenesses (the ε and ξ) associated with outcomes and measurements for each stopping time. CHH only use information on the covariances to identify the factor loadings.[53] In place of the information from the index generating choices that was used in the analysis of Section 2.5, in this section, because we are using limit sets that fix treatment times, it is necessary to use measurements to produce the

[52] The average treatment effects are identified using only the marginal distributions.

[53] They also assume a "triangular" structure on the matrix of factor loadings for their principal results. This structure assumes that there are two (or more) measurements or outcomes that depend only on θ_1; two (or more) measurements or outcomes that depend only on θ_1 and θ_2 and so forth. Use of covariance information limits the number of factors that can be nonparametrically identified. Thus for an outcome and measurement vector J of length N, $N > 2L + 1$

$$J = \underbrace{\begin{pmatrix} \alpha_{11} & 0 & 0 & \cdots & 0 \\ \alpha_{21} & 0 & 0 & \cdots & 0 \\ \alpha_{31} & \alpha_{32} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ \alpha_{N1} & \alpha_{N2} & \alpha_{N3} & \cdots & \alpha_{NL} \end{pmatrix}}_{N \times L} \begin{pmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \\ \vdots \\ \theta_L \end{pmatrix} + \begin{pmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \\ \vdots \\ \xi_N \end{pmatrix}$$

where the last block must have three or more elements and each of the $N - 3$ preceding rows has at least a block of two rows with the same pattern of zeros, the column vector of the ξ_i has mutually independent elements and is independent of the θ_i, and the θ_i are mutually independent. See Anderson and Rubin (1956). CHH also establish identification of a nontriangular structure.

factor loadings generating outcomes across treatment times.[54] Navarro (2004a) shows that if the distribution of the factors is not symmetric, we can uniquely identify the factor structure without any measurements if we gain additional information from the higher order moments beyond the covariances used in section 2.5. See also Bonhomme and Robin (2004).[55]

Notice that measurements, M, are not needed to prove Theorem 4. Notice further that,

[54] Without the measurements, using only covariance information on outcomes it is not possible to identify the sign of the factor loadings across systems of outcomes associated with different stopping times unless the conditions specified in Navarro (2004a) are satisfied. See Cunha, Heckman, and Navarro (2005e) or Carneiro, Hansen, and Heckman (2001) for intuitive discussions of these conditions. See also the survey in Heckman, Lochner, and Todd (2006).

[55] The following example that builds on the analysis of Section 2.5 illustrates Navarro's results and the related results by Bonhomme and Robin (2004). Assume a one factor model and two systems associated with two stopping times identified in limit sets. We use three outcomes. The first subscript defines the system used:

$$\begin{aligned} Y_{0,1} &= \alpha_{0,1}\theta + \varepsilon_{0,1} \qquad Y_{1,1} = \alpha_{1,1}\theta + \varepsilon_{1,1} \\ Y_{0,2} &= \alpha_{0,2}\theta + \varepsilon_{0,2} \qquad Y_{1,2} = \alpha_{1,2}\theta + \varepsilon_{1,2} \\ Y_{0,3} &= \alpha_{0,3}\theta + \varepsilon_{0,3} \qquad Y_{1,3} = \alpha_{1,3}\theta + \varepsilon_{1,3} \end{aligned}$$

where $\theta \perp\!\!\!\perp (\varepsilon_{0,1}, \varepsilon_{0,2}, \varepsilon_{0,3}, \varepsilon_{1,1}, \varepsilon_{1,2}, \varepsilon_{1,3})$ and the $\varepsilon_{i,j}$ are mutually independent and mean zero with finite variances. θ has mean zero and a finite variance. We observe data on the $(Y_{0,1}, Y_{0,2}, Y_{0,3})$ system or the $(Y_{1,1}, Y_{1,2}, Y_{1,3})$ system but not both.

Suppose we normalize $\alpha_{0,1} = 1$. Using the analysis of Section 2.5, we can identify $\alpha_{0,2}, \alpha_{0,3}$ and the distributions of $\theta, \varepsilon_{0,1}, \varepsilon_{0,2}, \varepsilon_{0,3}$ nonparametrically, from the first system. From the second system, we can identify

$$\begin{aligned} Cov(Y_{1,1}, Y_{1,2}) &= \alpha_{1,1}\alpha_{1,2}\sigma_\theta^2 \\ Cov(Y_{1,1}, Y_{1,3}) &= \alpha_{1,1}\alpha_{1,3}\sigma_\theta^2 \\ Cov(Y_{1,2}, Y_{1,3}) &= \alpha_{1,2}\alpha_{1,3}\sigma_\theta^2 \end{aligned}$$

Then, assuming $\alpha_{1,1} \neq 0, \alpha_{1,2} \neq 0, \alpha_{1,3} \neq 0$ and $\sigma_\theta^2 > 0$, we can identify $\frac{Cov(Y_{1,1},Y_{1,2})}{Cov(Y_{1,1},Y_{1,3})} = \frac{\alpha_{1,2}}{\alpha_{1,3}}$ so $\alpha_{1,2} = \frac{Cov(Y_{1,1},Y_{1,2})}{Cov(Y_{1,1},Y_{1,3})}\alpha_{1,3}$. We also can obtain $Cov(Y_{1,2}, Y_{1,3}) = \frac{Cov(Y_{1,1},Y_{1,2})}{Cov(Y_{1,1},Y_{1,3})}\alpha_{1,3}^2\sigma_\theta^2$. Thus we obtain

$$\alpha_{1,3}^2 = \frac{Cov(Y_{1,2}, Y_{1,3})Cov(Y_{1,1}Y_{1,3})}{Cov(Y_{1,1}, Y_{1,2})\sigma_\theta^2}.$$

The sign of $\alpha_{1,3}$ is not determined.

If, however, we use the assumption that θ is non-normal and $E(\theta^3) \neq 0$, we can form

$$E(Y_{1,1}Y_{1,3}^2) = \alpha_{1,1}\alpha_{1,3}^2 E(\theta^3)$$

and hence we can solve for $\alpha_{1,1}$ from

$$\alpha_{1,1} = \frac{E(Y_{1,1}Y_{1,3}^2)}{\alpha_{1,3}^2 E(\theta^3)}$$

where we know all of the ingredients on the right hand side. Thus we can identify $\alpha_{1,2}, \alpha_{1,3}$ and hence can form the joint distribution of $(Y_{0,1}, Y_{0,2}, Y_{0,3}, Y_{1,1}, Y_{1,2}, Y_{1,3})$. Navarro shows that we need only one measurement per factor so one can relax the bound $N > 2L + 1$. There is related work by Bonhomme and Robin (2004).

while there are formal similarities to the duration model for time to treatment developed in Section 2, there are important differences that arise because the model of this section is forward-looking. For example, the index $I\left(a, t\left(a\right), \mathcal{I}_{t(a)}\right)$ is a function of expected future outcomes. The proof strategy used in Section 2 is applied in reverse order.

We now establish identifiability of the parameters of the last choice index, including the parameters of the cost function. We then proceed to identify the next to last index and proceed backward to the initial stage choice index. We start by analyzing the last transition (from $\bar{T}-1$ to $\bar{T}$). Notice that once an individual is at $\bar{T}$, his remaining lifetime value is no longer a function of cost (and hence Z) since no further transitions are possible. The temporal structure of the finite horizon decision problem produces natural exclusion restrictions and we exploit it. We now prove the following theorem which demonstrates this point.

Theorem 5. *Assume that conditions (i)–(iv) of Theorem 4 hold. In particular, one implication of condition (iv) of Theorem 4 is especially important in this proof:*

(*) $Supp\left(\Phi\left(\bar{T}-1, X, Z\left(\bar{T}-1\right)\right)\right) \supseteq Supp\left(-\Upsilon\left(\bar{T}-1, X\right)\right)$.

We assume that $\Phi\left(\bar{T}-1, X, Z\left(\bar{T}-1\right)\right)$ *belongs to the class of Matzkin functions, and that* r *is known. Then, the mean cost function* $\Phi(\bar{T}-1, X, Z\left(\bar{T}-1\right))$*, the marginal distribution of* $\Upsilon\left(\bar{T}-1, X\right)$*, the factor loadings* $\lambda_{\bar{T}-1}$ *and the distribution of* $\xi(\bar{T}-1)$ *are identified for all* X*. If we specify the* $\Phi\left(\bar{T}-1, X, Z\left(\bar{T}-1\right)\right)$ *only up to scale, we identify the cost function and the marginal distribution of* $\Upsilon\left(\bar{T}-1, X\right)$ *up to the scale as well as the distribution of* $\xi(\bar{T}-1)$*.*

Proof. As a consequence of assumptions (*iii*) and (*iv*) of Theorem 4, a limit set $\tilde{\mathcal{Z}}$ exists such that $\lim_{X,Z\to\tilde{\mathcal{Z}}} \Pr(T > \bar{T}-2 \mid X, Z) = 1$. One way to obtain the limit set is to assume that for all $t\left(a\right)$ there is at least one continuous variable $Z_j\left(t\left(a\right)\right)$ that is not contained in any other $Z\left(t\left(a'\right)\right)$ $(a' \neq a)$ (where subscripts denote components of $Z(t(a))$) or in X, that $\Phi\left(t\left(a\right), X, Z\left(t\left(a\right)\right)\right)$ is monotonic in $Z_j\left(t\left(a\right)\right)$, that there are no restrictions on the supports, and that variation in $Z_j(t(a))$ traces out the full support of $\Upsilon(X, Z)$ to satisfy

(iv) of Theorem 4. However, we can satisfy this requirement without having a conventional exclusion.

Recall that an individual, conditional on having reached $\bar{T}-1$ with probability one in the limit set $\tilde{\mathcal{Z}}$ and conditional on $X=\tilde{x}$ and $Z(\bar{T}-1)=z(\bar{T}-1)$, will stop at stage $\bar{T}-1$ if $\Phi\left(\bar{T}-1,\tilde{x},z\left(\bar{T}-1\right)\right)+\Upsilon\left(\bar{T}-1,\tilde{x}\right)>0$. Under assumption (iii) of Theorem 4, we can freely vary $\Phi(\bar{T}-1,\tilde{x},z\left(\bar{T}-1\right))$ by varying $z(\bar{T}-1)$ while keeping $X=\tilde{x}$ fixed. Alternatively, we could fix $\mu\left(X\right)=k$ and still be able to vary Φ without having to fix the entire X vector since $\Upsilon\left(\bar{T}-1,\tilde{x}\right)$ only depends on X through the effect of mean earnings on the value functions. In this way we would not require that some elements of Z be different from elements of X. Observe that $\Upsilon\left(\bar{T}-1,\tilde{x}\right)$ is a random variable that is statistically independent of $Z(\bar{T}-1)$ given $X=\tilde{x}$ (or $\mu\left(X\right)=k$). Since we can freely vary $\Phi\left(\bar{T}-1,\tilde{x},z\left(\bar{T}-1\right)\right)$ in the limit set, conditional on $X=\tilde{x}$, we can use standard proofs for identification in a binary choice model. If $\Phi(\bar{T}-1,X,Z\left(\bar{T}-1\right))$ is in the Matzkin class, we can identify $\Phi\left(\bar{T}-1,\tilde{x},z\left(\bar{T}-1\right)\right)$ over its support for $X=\tilde{x}$ and the distribution of $\Upsilon\left(\bar{T}-1,\tilde{x}\right)$ for a given $X=\tilde{x}$.

In the limit set, and conditional on $X=\tilde{x}$, $Z(\bar{T}-1)=z(\bar{T}-1)$, from the data, we can form

$$\begin{aligned}&\Pr\left(M<m(\tilde{x}),Y(\bar{T})<y\left(\bar{T},\tilde{x}\right)\mid T=\bar{T},X=\tilde{x},Z\left(\bar{T}-1\right)=z\left(\bar{T}-1\right)\right)\\&\times\Pr\left(T=\bar{T}\mid X=\tilde{x},Z\left(\bar{T}-1\right)=z\left(\bar{T}-1\right)\right).\end{aligned}$$

Varying $y(\bar{T},\tilde{x})$, $m(\tilde{x})$ and $-\Phi\left(\bar{T}-1,\tilde{x},z\left(\bar{T}-1\right)\right)$, and adjusting for $\mu_M(\tilde{x})$ and $\mu(\tilde{x})$ we can identify the distribution of the unobservables $\left(U_M,U\left(\bar{T}\right),\Upsilon\left(\bar{T}-1,\tilde{x}\right)\right)$ at known points of evaluation:

$$F_{U_M,U(\bar{T}),\Upsilon\left(\bar{T}-1,\tilde{x}\right)}\left(\begin{array}{c}m-\mu_M\left(\tilde{x}\right),y-\mu\left(\bar{T},\tilde{x}\right),\\-\Phi\left(\bar{T}-1,\tilde{x},z\left(\bar{T}-1\right)\right)\end{array}\middle|\ X=\tilde{x},Z\left(\bar{T}-1\right)=z\left(\bar{T}-1\right)\right).$$

Notice that we require that both $Y(\bar{T})$ and M be continuous random variables so that we can trace the distribution conditional on $X = \tilde{x}$ by varying $y(\bar{T}, \tilde{x})$ and $m(\tilde{x})$.[56]

Up to this point in the proof we have not invoked a factor structure and it is not needed. If we invoke a factor structure from the distribution of $\left(U_M, U\left(\bar{T}\right), \Upsilon\left(\bar{T}-1, \tilde{x}\right)\right)$ we can identify the factor loadings on the θ in the cost functions. To show this, suppose that the unobservables associated with measurements M_1, a subvector of M, only depend on θ_1 and not the other factors generating the model. Since we have identified the joint distribution of $\left(U_M, U\left(\bar{T}\right), \Upsilon\left(\bar{T}-1, \tilde{x}\right)\right)$ conditional on $X = \tilde{x}, Z(\bar{T}-1) = z(\bar{T}-1)$, we can construct the left hand side of

$$Cov\left(U_{1,M}, \Upsilon\left(\bar{T}-1, \tilde{x}\right) \mid X = \tilde{x}, Z(\bar{T}-1) = z(\bar{T}-1)\right) =$$

$$Cov\left(U_{1,M}, \left[\begin{array}{r} R\left(\bar{T}-1, \bar{T}-1, \tilde{x}\right) - Y\left(\bar{T}-1, \bar{T}-1, 0, \tilde{x}\right) \\ -\frac{1}{1+r} E\left(R\left(\bar{T}, \bar{T}, \tilde{x}\right) \mid \mathcal{I}_{\bar{T}-1}\right) \end{array}\right] \middle| \begin{array}{l} X = \tilde{x}, \\ Z(\bar{T}-1) = z(\bar{T}-1) \end{array}\right)$$

$$+\alpha_{1,1,M}\lambda_{1,\bar{T}-1}\sigma^2_{\theta_1},$$

where on the right hand side of the equation $\lambda_{1,\bar{T}-1}$ is the coefficient on θ_1 in the $(\bar{T}-1)^{st}$ cost function. The left hand side of this expression, the first term on the right hand side, $\alpha_{1,1,M}$ and $\sigma^2_{\theta_1}$ are known from Theorem 4 after applying factor analysis to outcomes and measurements. (This assumes that either the conditions in CHH or Navarro, 2004a, are met.) We can use this covariance to identify $\lambda_{1,\bar{T}-1}$, since we know $\alpha_{1,1,M}$ and $\sigma^2_{\theta_1}$ from a factor analysis of the measurement system. Proceeding sequentially, taking covariances of the choice index with equations that depend on additional elements of θ, we identify all of the loadings $\lambda_{\bar{T}-1}$ associated with the cost function under the conditions on the factors specified in CHH or Navarro (2004a). This requires additional measurements M that depend on the factors in the cost function. In addition, this analysis assumes a triangular factor loading

[56] We only need some components of M to be continuous. See CHH or the analysis in Appendix D.

matrix as previously discussed.[57] Once knowledge of the $\lambda_{\bar{T}-1}$ is secured, we can identify the distribution of $\xi\left(\bar{T}-1\right)$ by using deconvolution applied to the distribution of $\Upsilon\left(\bar{T}-1,\tilde{x}\right)$, which is nonparametrically identified. $\Upsilon\left(\bar{T}-1,X\right)$ can be represented as

$$\begin{aligned}\Upsilon\left(\bar{T}-1,X\right) &= \xi\left(\bar{T}-1\right)+\theta\lambda_{\bar{T}-1}+R\left(\bar{T}-1,\bar{T}-1,X\right) \\ &\quad -Y\left(\bar{T}-1,\bar{T}-1,0,X\right)-\frac{1}{1+r}E\left(R\left(\bar{T},\bar{T},X\right)\mid\mathcal{I}_{\bar{T}-1}\right),\end{aligned} \tag{5}$$

where $X=\tilde{x}$, $Z(\bar{T}-1)=z(\bar{T}-1)$ are contained in the agent's information set at $\bar{T}-1$, $\mathcal{I}_{\bar{T}-1}$. We identify the Y functions and their distribution for all ages for each t as a consequence of Theorem 4. Thus we can construct the R functions and their distribution which only depend on the Y functions and their distribution. We know the factor loadings and the distribution of the factors (θ). Hence we know the distribution of $R(\bar{T}-1,\bar{T}-1,\tilde{x})-Y(\bar{T}-1,\bar{T}-1,0,\tilde{x})-\frac{1}{1+r}E\left(R\left(\bar{T},\bar{T},X\right)\mid\mathcal{I}_{\bar{T}-1}\right)$. Therefore we know the distribution of the sum of the terms on the right hand side after $\xi\left(\bar{T}-1\right)$ in the expression $\Upsilon\left(\bar{T}-1,X\right)$. By assumption, $\xi\left(\bar{T}-1\right)$ is independent of the remaining terms on the right hand side. Finally, we can vary $X=x$ to identify the $\Phi\left(\bar{T}-1,x,z\left(\bar{T}-1\right)\right)$ for all $X=x$ up to the scale of $\Upsilon\left(\bar{T}-1,x\right)$. We can construct all of the components of the distribution of $\Upsilon\left(\bar{T}-1,x\right)$ and the joint distribution of any subcomponent. Hence we also know the scale of $\Upsilon\left(\bar{T}-1,X\right)$ for all $X=x$. ■

Observe that we do not need any measurements M to identify the joint distribution of $U\left(\bar{T}\right),\Upsilon\left(\bar{T}-1,\bar{x}\right)$ or the mean of the cost function $\Phi\left(\bar{T}-1,\tilde{x},z\left(\bar{T}-1\right)\right)$. The measurements are only used to recover the distribution of the unobservables in the cost function and the associated factor loadings. Thus we can identify the discrete choice model and the associated outcome without using any measurements.

Theorem 5 establishes conditions under which we can identify all of the elements of the cost function for the last transition. We can determine the scale if one element of cost is

[57] This proof can be modified to accommodate other factor structure assumptions but we do not do so here.

known to the econometrician (*e.g.* tuition). This corresponds to a special case of the Matzkin functions. This analysis is predicated on a particular information set. A component of the information set used in the proof of Theorem 5 is that θ_1 is known to the agent at $\bar{T}-1$. If it is not, then $\lambda_{1,\bar{T}-1}=0$ and our proof simplifies. Alternative specifications of the information set produce different distributions of $\Upsilon\left(\bar{T}-1,X\right)$ and more generally $\Upsilon\left(X\right)$.

The proof assumes a known interest rate r. This assumption simplifies the proof but is not essential to it. To see how r is identified, note that under assumption (ii) of Theorem 5, the terminal values $R\left(\bar{T}-1,\bar{T}-1,X\right)$ and $R\left(\bar{T},\bar{T},X\right)$ depend on X only through the means of the $Y\left(t,1,X\right)$ equations. See equation (3) for the explicit representation and equation (4) for the definition of the R terms. Under our assumptions about the information known to the agent, (including assumptions (I-1) and (I-2)), and because of the independence produced from assumption (ii), $E\left(R\left(\bar{T},\bar{T},X\right)\mid\mathcal{I}_{\bar{T}-1}\right)$ also depends on X only through the mean functions $\mu\left(a,t\left(a\right),1,X\right)$ in equation (3).[58]

If we adjoin to the assumptions invoked in Theorem 5, the assumption that

Supplementary Assumption (**) to Theorem 5: *$\mu\left(\bar{T}-1,\bar{T}-1,1,X\right)$ and $\mu\left(\bar{T},\bar{T},1,X\right)$ are continuous and differentiable in at least one argument of X,*

we can use the index property of the choice model to compute

$$\frac{\dfrac{\partial\Pr\left(T=\bar{T}\mid X,Z\right)}{\partial\mu\left(\bar{T}-1,\bar{T}-1,1,X\right)}}{\dfrac{\partial\Pr\left(T=\bar{T}\mid X,Z\right)}{\partial\mu\left(\bar{T},\bar{T},1,X\right)}}=1+r \tag{6}$$

because we can freely vary the mean functions generating $R\left(\bar{T}-1,\bar{T}-1,X\right)$ and $R\left(\bar{T},\bar{T},X\right)$ under assumption (iii) of Theorem 4, and the derivatives exist because we assume that the random variables generating the unobservables in (5) are absolutely continuous with respect to Lebesgue measure. Clearly we can use other combinations of the mean functions generating $R\left(\bar{T}-1,\bar{T}-1,X\right)$ and $R\left(\bar{T},\bar{T},X\right)$ to identify r, provided a version of assumption

[58] We know $\mu\left(a,t\left(a\right),1,X\right)$ as a result of Theorem 4.

(**) holds for the selected mean functions. Observe that the choice of a scale function for the Matzkin class is irrelevant since the scale cancels. Formula (6) for the Matzkin class is a version of Powell, Stock, and Stoker (1989) or Horowitz (1998).

If we only specify the Matzkin class of functions up to scale, this theorem is not strong enough to identify the cost functions for the preceding transitions even up to scale. In the transitions before $\bar{T}-1$, costs appear in the final reward functions. Thus the choice index for transition $\bar{T}-2$ is

$$\begin{aligned} I\left(\bar{T}-2,\bar{T}-2,\mathcal{I}_{\bar{T}-2}\right) &= R\left(\bar{T}-2,\bar{T}-2,X\right)-K\left(\bar{T}-2,\bar{T}-2,\mathcal{I}_{\bar{T}-2}\right) \\ &= R\left(\bar{T}-2,\bar{T}-2,X\right)-Y\left(\bar{T}-2,\bar{T}-2,0,X\right) \\ &\quad +C\left(\bar{T}-2,X,Z\left(\bar{T}-2\right)\right) \\ &\quad -\frac{1}{1+r}E\left(V\left(\bar{T}-1,\bar{T}-1,\mathcal{I}_{\bar{T}-1}\right)\mid\mathcal{I}_{\bar{T}-2}\right) \end{aligned}$$

and

$$V\left(\bar{T}-1,\bar{T}-1,\mathcal{I}_{\bar{T}-1}\right)=\max\left\{R\left(\bar{T}-1,\bar{T}-1,X\right),\left(\begin{array}{l} Y\left(\bar{T}-1,\bar{T}-1,0,X\right) \\ -C\left(\bar{T}-1,X,Z\left(\bar{T}-1\right)\right) \\ +\frac{1}{1+r}E\left(R\left(\bar{T},\bar{T},X\right)\mid\mathcal{I}_{\bar{T}}\right)\end{array}\right)\right\}.$$

Knowledge of $C\left(\bar{T}-1,X,Z\left(\bar{T}-1\right)\right)$ measured in the same scale as $R\left(\bar{T}-2,\bar{T}-2,X\right)$ is required to form $V\left(\bar{T}-1,\bar{T}-1,\mathcal{I}_{\bar{T}-1}\right)$. The following theorem, which draws on Matzkin (1994), gives two conditions under which the unknown scale on the cost function can be determined, if it is not specified by assuming that $\Phi\left(\bar{T}-1,X,Z\left(\bar{T}-1\right)\right)$ is in the Matzkin class.

Theorem 6. *Assume either that:*

(i) It is possible to partition $X=\left(\hat{X},\tilde{X}\right)$ *so that the elements of* $\hat{X}$ *do not enter* $\Phi\left(\bar{T}-1,X,Z\left(\bar{T}-1\right)\right)$. *Furthermore, assume additive separability of the mean out-*

come function for $Y(\bar{T}-1,\bar{T}-1,j,X)$ *in terms of the two components:*

$$\mu\left(\bar{T}-1,\bar{T}-1,j,X\right)=n^{*}\left(\bar{T}-1,\bar{T}-1,j,\hat{X}\right)+n\left(\bar{T}-1,\bar{T}-1,j,\tilde{X}\right),\ j=0,1.$$

Alternatively, assume that

(ii) It is possible to partition $Z\left(\bar{T}-1\right)=\left(\hat{Z}\left(\bar{T}-1\right),\tilde{Z}\left(\bar{T}-1\right)\right)$ *and that the cost function has an additively separable component with a known coefficient:*

$$\Phi\left(\bar{T}-1,X,Z\left(\bar{T}-1\right)\right)=\phi\left(\bar{T}-1,X,\tilde{Z}\left(\bar{T}-1\right)\right)+\hat{Z}\left(\bar{T}-1\right)$$

so that $\hat{Z}\left(\bar{T}-1\right)$ *is measured in the same units as* $R\left(\bar{T}-2,\bar{T}-2,X\right)$*. This would be the case if, for example,* $\hat{Z}\left(\bar{T}-1\right)$ *measured direct costs of schooling (*e.g. *tuition in our schooling example).*

Then, if either (i) or (ii), or both hold, the scale of $\Upsilon\left(\bar{T}-1,X\right)$ *in Theorem 5 is identified.*

Proof. Part (ii) is immediate, because we set the scale of one coefficient and can use its identified coefficient in the choice equation to identify the scale of $\Upsilon\left(\bar{T}-1,X\right)$. (See Matzkin (1994)). Part (i) is also straightforward because we can determine $n^{*}\left(\bar{T}-1,\bar{T}-1,1,\tilde{x}\right)$ from the limit sets of the outcome equations and it also enters the choice equation and identifies the scale. ■

We next consider identification of the cost function for transition $\bar{T}-2$ under the assumption that we can identify the scale at $\bar{T}-1$. The distribution of $\Upsilon\left(\bar{T}-2,X,Z^{\bar{T}-1}\right)$ depends on X and $Z\left(\bar{T}-1\right)$ because all future returns and costs are in the value function. The key insight in our theorem is to note that the dependence on $Z\left(\bar{T}-1\right)$ is not general but operates through the function $\Phi\left(\bar{T}-1,X,Z\left(\bar{T}-1\right)\right)$ which was identified by the preceding argument.

Theorem 7. *Assume conditions (i)–(iv) of Theorem 4. Assume that* $\Phi\left(\bar{T}-1,X,Z\left(\bar{T}-1\right)\right)$ *is in the Matzkin class of functions. In particular, we assume that*

(*) $Supp\left(\Phi\left(\bar{T}-2, X, Z\left(\bar{T}-2\right)\right)\right) \supseteq Supp\left(-\Upsilon\left(\bar{T}-2, X, Z^{\bar{T}-1}\right)\right)$ *which follows from (iv) of Theorem 4*

and

(**) $Supp\left(\Phi\left(\bar{T}-2, X, Z\left(\bar{T}-2\right)\right), \Phi\left(\bar{T}-1, X, Z\left(\bar{T}-1\right)\right)\right) =$

$Supp\left(\Phi\left(\bar{T}-2, X, Z\left(\bar{T}-2\right)\right)\right) \times Supp(\Phi(\bar{T}-1, X, Z\left(\bar{T}-1\right)))$ *which follows from condition (iii) of Theorem 4 applied element by element.*

In addition to assumptions (i)-(iv) of Theorem 4, assume that

(***) The conditions of Theorem 6 apply so that we can identify the scale of the cost function in the last transition, $\bar{T}-1$.

Then, $\Phi\left(\bar{T}-2, X, Z\left(\bar{T}-2\right)\right)$, *the marginal distribution of* $\Upsilon\left(\bar{T}-2, X, Z^{\bar{T}-1}\right)$, *the factor loadings* $\lambda_{\bar{T}-2}$, *and the distribution of* $\xi(\bar{T}-2)$ *are identified for all* $X, Z^{\bar{T}-2}$. *Alternatively if we specify the Matzkin class of functions up to scale we identify* $\Phi\left(\bar{T}-2, X, Z\left(\bar{T}-2\right)\right)$ *and the distribution of* $\Upsilon\left(\bar{T}-2, X, Z^{\bar{T}-1}\right)$ *up to scale.*

Proof. From Theorem 4, a limit set exists such that $\Pr(T > \bar{T}-3 \mid X=x, Z=z)=1$. Consider, in this limit set,

$$\begin{aligned} & \Pr\left(T=\bar{T}-2 \mid X=x, Z(\bar{T}-1)=z(\bar{T}-1), Z(\bar{T}-2)=z(\bar{T}-2)\right) \\ = \; & \Pr\left(\Phi(\bar{T}-2, x, z(\bar{T}-2))+\Upsilon\left(\bar{T}-2, x, z(\bar{T}-1)\right) \geq 0\right). \end{aligned}$$

Observe that $\Upsilon\left(\bar{T}-2, x, z(\bar{T}-1)\right)$ depends on $z(\bar{T}-1)$ only through $\Phi(\bar{T}-1, x, z(\bar{T}-1))$. As a consequence, we can express the preceding probability as

$$\Pr(\Phi(\bar{T}-2, x, z(\bar{T}-2))+\Upsilon^*\left(\bar{T}-1, X, \Phi(\bar{T}-1, x, z(\bar{T}-1))\right))>0$$

where $\Upsilon^*\left(\bar{T}-1, x, \Phi(\bar{T}-1, x, z(\bar{T}-1))\right)$ shows the explicit dependence of $\Upsilon\left(\bar{T}-2, x, z(\bar{T}-1)\right)$ on the mean cost function $\Phi(\bar{T}-1, x, z(\bar{T}-1))$. From assumption *(iii)* of Theorem 4, we

can condition on $\Phi(\bar{T}-1, x, z(\bar{T}-1)) = \varphi$ and still be able to vary $\Phi(\bar{T}-2, x, z(\bar{T}-2))$ freely. Therefore we can trace out the distribution of $\Upsilon\left(\bar{T}-2, x, z(\bar{T}-1)\right)$ analogous to the way we traced out the distribution of $\Upsilon\left(\bar{T}-1, x\right)$ in the proof of Theorem 5, and we can construct the joint distribution of U_M, $U(\bar{T}-2)$, $\Upsilon\left(\bar{T}-2, x, z(\bar{T}-1)\right)$. We can mimic the proof of Theorem 5 and identify $\lambda_{\bar{T}-2}$ and the distribution of $\xi(\bar{T}-2)$. We can do this for all $Z(\bar{T}-2) = z(\bar{T}-2)$, $Z(\bar{T}-1) = z(\bar{T}-1)$ and $X = x$. As in the proof of Theorem 5, instead of conditioning on X we can also condition on $\mu(X) = k$ and we can still vary for $\Phi\left(\bar{T}-2, x, Z\left(\bar{T}-1\right)\right)$ without having to fix the entire X vector. In this way we do not require some elements of Z to be different from X.If we specify the Matzkin class only up to scale the proof only goes through for $\Phi\left(\bar{T}-2, x, Z\left(\bar{T}-1\right)\right)$ up to the unknown scale and the distribution of the unobservables up to scale. ■

The easiest way to satisfy the conditions of Theorem 7 is to assume access to $Z(t)$, $t = 1, \ldots, \bar{T}-1$, that are mutually statistically independent of each other. This is far stronger than what is required to secure identification. We can allow the $Z(t)$ to be dependent but we need to rule out any degeneracy in the joint distribution of $Z(t)$, $t = 1, \ldots, \bar{T}$. $Z(t)$ variables with these properties would arise if there are stopping-time-specific cost variables (*e.g.* college tuition for college; school fees for secondary levels, etc.). However, Theorem 7 would still apply if the *same* Z variables appear in each stopping-time-specific cost function, provided that we satisfy the generalization of condition (*iii*) of Theorem 4. Theorem 7 is a generalization of Corollary 1 of Section 2 that applies to an explicitly formulated, forward-looking model.[59] When the same Z appears in each cost function, it is required that the curvature of the mean cost functions differ across stopping times in order to satisfy the condition. It is necessary that Z be a vector. If Z is scalar, condition (*iii*) of Theorem 4 fails. We need to modify Theorem 6 to identify the absolute scale of the cost function at stage $T-2$.

[59] Corollary 1 applies only to the index functions as they enter the limits of the integrals generating the expressions. Theorem 7 generalizes this result to include dependence of the distributions of the generated random variables on the index functions of the model.

Proceeding sequentially across stopping times with suitably modified conditions *(i)–(iv)* in Theorem 4, enables us to identify all of the cost functions at all stages of the process provided that we modify Theorem 6 appropriately. This allows us, for each stopping time, to identify private valuations (costs) and separate them from objective outcomes. Thus, in the context of models of health economics, we can separate outcomes of a treatment from the psychic costs of taking it at a particular time.[60]

In this model, analysts can distinguish period by period *ex ante* expected returns from *ex post* realizations by applying the analysis of Cunha, Heckman, and Navarro (2005e) and Navarro (2004b). Because we can link choices to outcomes through the factor structure assumption, we can also distinguish *ex ante* preference or cost parameters from their *ex post* realizations. *Ex ante*, agents may not know θ. *Ex post*, they do. All of the information about future rewards and returns is embodied in the information set $\mathcal{I}_a$. Unless the time of treatment is known with perfect certainty, it cannot cause outcomes prior to its realization. Thus in an environment of uncertainty we rule out the possibility that the future can cause the past—a possibility that is not ruled out in the reduced form models of Section 2, except by imposing it directly onto the parameters of the model.

Our analysis is predicated on specification of the agent's information sets which should be carefully distinguished from the econometrician's. Cunha, Heckman, and Navarro (2005e) and Navarro (2004b) present methods for determining which components of future outcomes are in the information sets of agents at each age, $\mathcal{I}_a$. If they are unknown to the agent at age a, under rational expectations, agents form their value functions used to make schooling choices by integrating out the unknown components using the distributions in their information sets. Components that are known to the agent are treated as constants by the individual in forming the value function but as unknown variables by the econometrician and their distribution is estimated. The true information set of the agent is determined from the set of possible specifications of the information sets of agents by picking the specification that best

[60] In the analysis of CHH and Cunha, Heckman, and Navarro (2005a,b,c,e), psychic costs of schooling are distinguished from monetary returns.

fits the data on choices and outcomes penalizing for parameter estimation. Heuristically, if neither the agent nor the econometrician knows a variable, the econometrician identifies the determinants of the distribution of the unknown variables that is used by the agent to form expectations. If the agent knows some variables, but the econometrician does not, the econometrician seeks to identify the distribution of the variables, but the agent treats the variables as known constants.

We can identify all of the treatment parameters including pairwise *ATE*, the marginal treatment effect *MTE* for each transition (obtained by finding mean outcomes for individuals indifferent between transitions), all of the treatment on the treated and treatment on the untreated parameters and the population distribution of treatment effects by applying the analysis of CHH and Cunha, Heckman, and Navarro (2005e) to this model. See also the discussion in appendix B. Our analysis can easily be generalized to cover the case where there are vectors of contemporaneous outcome measures for different stopping times and ages, building on the analysis of Appendix D modified to suit this more precisely formulated choice model. We next discuss how to implement the limit set strategy.

3.2 Implementing the limit set strategy and checking for identification

Under our assumptions, the limit sets used in the theorems in this paper are obtained by finding subsets of the data that make the probabilities of each stopping time T arbitrarily close to 1. In any sample, we can check whether such subsets exist or are very thin, since we can nonparametrically compute $\Pr(T = t \mid Z = z, X = x)$. Figure 2, taken from the research of Heckman, Stixrud, and Urzua (2004), shows the result of such an analysis.[61] It plots the sample distribution of probabilities of final schooling attainment (at age 30) for males over all subsets of (X, Z) in the data. In the sample, one cannot find any subset with mass in the probabilities near 1 for any final schooling choice. Thus in their sample, the required limit

[61] See also analysis of Heckman and Navarro (2006).

sets.

One can argue that this problem will vanish in large samples. That is an assumption that cannot be checked with data. Alternatively, one can argue that we obtain identification of the distributions over a subset and develop bounds on the model (see Manski, 2003). Another alternative is to assume that the partially identified distributions are real analytic and continue them over the missing support using analytic continuation.[62]

Our limit set arguments identify outcome distributions for values of choice probabilities that become big or small. They have a close resemblance to the assumption used in the recent nonparametric structural literature (see Matzkin, 1994, 2003) that the econometrician knows the function sought to be identified at some point, or points, of evaluation. In our context, the function is an outcome distribution. That literature is unclear about how to select the points of evaluation whereas our analysis provides guidance in terms of large or small values of the probability of selection into states. We next turn to a comparison of the reduced form and structural models analyzed in this paper.

3.3 Comparing Reduced Form and Structural Models

The reduced form model analyzed in Section 2 is typical of many reduced form statistical approaches within which it is difficult to make important conceptual distinctions. Because the choice equation is not modeled explicitly, it is hard to use such frameworks to analyze the decision makers' expectations, costs of treatment, the arrival of information, the content of agent information sets and the consequences of the arrival of information for decisions regarding time to treatment as well as outcomes. In particular, it is difficult to distinguish *ex post* from *ex ante* valuations of outcomes. Cunha, Heckman, and Navarro (2005e), Navarro (2004b) and Heckman and Navarro (2006) present analyses that distinguish *ex ante* anticipations from *ex post* realizations.[63] In reduced form models, it is difficult to make the distinction between private evaluations and preferences (*e.g.* "costs" as defined in this

[62] Heckman and Singer (1984) discuss this strategy.

[63] See the summary of this literature in Heckman, Lochner, and Todd (2006).

section) from objective outcomes (the Y variables).

Statistical and reduced form econometric approaches to analyzing dynamic counterfactuals appeal to uncertainty to motivate the stochastic structure of models. They do not explicitly characterize how agents respond to uncertainty or make treatment choices based on the arrival of new information (see Robins, 1989, 1997, Lok, 2001, Gill and Robins, 2001, Abbring and Van Den Berg, 2003, and Van der Laan and Robins, 2003). In addition, as noted in section 2, in the reduced form models it is in principle possible to identify treatment effects where the future treatment time causes the past. Abbring and Van Den Berg (2003), Gill and Robins (2001) and Lok (2001) rule this out by imposing restrictions on the statistical treatment effect model.[64] The structural approach presented in this paper allows for a clear treatment of the arrival of information, agent expectations, and the effects of new information on choice and its consequences. In an environment of imperfect certainty about the future, it rules out the future causing the past once the effects of agent information sets are controlled for.

The structural model developed in this paper allows agents to learn about new factors (components of θ) as they proceed sequentially through their life cycles. It also allows agents to learn about other components of the model (see Cunha, Heckman, and Navarro, 2005e). Agent anticipations of when they will stop and the consequences of alternative stopping times can be revised sequentially. Their anticipated payoffs and stopping times are sequentially revised as new information becomes available. The mechanism by which agents revise their anticipations is modeled and identified. See Cunha, Heckman, and Navarro (2005a,b,c,e) for further discussion of these issues and Heckman, Lochner, and Todd (2006) for a partial survey of recent developments in the literature.

The clearest interpretation of the models in the statistical literature on dynamic treatment effects is as *ex post* selection-corrected analyses of distributions of events that have occurred. In a model of perfect certainty, where *ex post* and *ex ante* choices and outcomes

[64] This is the "nonanticipating" assumption of Abbring and Van Den Berg (2003).

are identical, the reduced form approach can be interpreted as a good approximation to a clearly specified choice model. In a more general analysis with information arrival and agent updating of information sets, the nature of the reduced form approximation is less clear cut. Thus it is unclear what agent decision-making processes and information arrival assumptions justify the conditional sequential randomization assumptions widely used in the dynamic treatment effect literature (see, *e.g.* Robins, 1989, 1997; Gill and Robins, 2001; Lok, 2001; Van der Laan and Robins, 2003; Lechner and Miquel, 2002) which are also used in branches of the dynamic discrete choice literature (see both Rust, 1987, and the survey in Rust, 1994). Reduced form approaches are not clear about the source of the unobservables and their relationship with conditioning variables. In reduced form analyses, the specification of the stochastic structure of the unobservables and the relationship of the unobservables to the observables is *ad hoc*. In the structural analysis, this specification emerges as part of the analysis, as our discussion of the stochastic properties of the unobservables presented in the preceding section makes clear.

The incompleteness intrinsic to reduced form models is illustrated in the analysis of Abbring and Van Den Berg (2003). They present an innovative and technically rigorous reduced form continuous time model of time to treatment where the treatment outcome is itself a continuous time duration. As Corollary 2 in Section 2.4 demonstrates, we can produce a discrete time counterpart to their model where the unobservables generating outcomes and the time to treatment equation and the relationship between the two sets of unobservables can be clearly modeled.

In their model, and in the reduced form models of Section 2, it is difficult to specify or determine what is in the agent's information set, how information is revised and the consequences of information revision for choices. They obtain their intuitively plausible "nonanticipation condition"—that the time of treatment does not affect pretreatment outcomes—by assuming that, conditional on time-invariant variables (both observed and unobserved by the econometrician), the pretreatment outcomes associated with two different treatment times

are the same up to and prior to the realization of the smaller of the two treatment times. Their condition rules out the possibility that the future can cause the past but at the price of assuming no learning about variables (observable and unobservable) that affect expectations of future outcomes and the choice of time to treatment after the process begins.

In our model, their assumption translates into the requirement that, conditional on initial observables and unobservables, the distribution of earnings while in high school is the same for those who become college graduates as it is for high school graduates who stop at that level of schooling. This assumption rules out any learning about ability, tuition costs, and the like, that can occur after the start of the process. We specify and identify different $Y(t, 0, X)$ processes for each information set. Agents with different expectations and agents with information sets that are revised over the courses of their life cycles may have different pre-treatment earnings and other outcome distributions. Using a well-posed economic model, we do not need to rule out learning in the structural model of Section 3 and we can still rule out the possibility that the future can cause the past. At each age $a = t(a)$ in the schooling process, agents update their information sets $\mathcal{I}_a = \mathcal{I}_{t(a)}$ and form new expectations about future outcomes. The mechanism for doing so is specified in the first part of this section. The reduced form treatment approach is incomplete in the sense of not providing a formal updating mechanism. Such updating is implicit in the conditioning sets that are sequentially updated (see, *e.g.* Gill and Robins, 2001; Lok, 2001).

Our analysis of both structural and reduced form models relies heavily on limit set arguments. They enable us to solve the selection problem in limit sets. The dynamic matching models of Gill and Robins (2001) and Lok (2001) solve the selection problem by invoking recursive conditional independence assumptions. In the context of our models, they assume that the econometrician knows the θ or can eliminate the θ by conditioning on a suitable set of variables. Our analysis entertains the possibility that analysts know substantially less than the agents they study. It allows for some of the variables that would make matching valid to be unobservable. Versions of recursive conditional independence assumptions are

also used in the dynamic discrete choice literature (see the survey in Rust, 1994). Our factor models allow us to construct the joint distribution of outcomes across stopping times. This feature is missing from the statistical treatment effect literature.

Both the structural and reduced form models share the property that it is possible to generate counterfactual treatment histories that are ruled out by a stopping time model. The index structure used to generate the model allows limits to be switched in the integrals based on latent variables — what we called the $B - D$ problem in Section 2.4. This feature is a consequence of the incomplete specification of both classes of models. We have not derived either reduced form or structural stopping models from a more basic model with the possibility of return from dropout states but which nonetheless exhibit the stopping time property. Our identification strategy in this paper relies on the nonrecurrent nature of treatment. We leave the task of formulating and identifying a general recurrent state version of the model for another occasion.[65]

4 Relationship of Our Work to Previous Work

Rust (1994) presents a widely cited nonparametric nonidentification theorem for dynamic discrete choice models. It is important to note the restrictive nature of his results. He analyzes a recurrent state infinite horizon model in a stationary environment. He does not exploit choice-specific outcome information nor does he use any exclusion restrictions or cross outcome-choice restrictions. He places no restrictions on period-specific utility functions such as concavity or linearity.

Magnac and Thesmar (2002) present an extended comment on Rust's analysis including positive results for identification when the econometrician knows the distributions of unobservables, assumes that unobservables enter period-specific utility functions in an additively separable way and is willing to specify functional forms of utility functions or other ingredi-

[65] Our identification strategy of using limit sets can be applied to the nonrecurrent model provided that we confined subsets of (X, Z) such that in those subsets the probability of recurrence is zero. See Heckman, Urzua, and Yates (2005).

ents of the model, as do Pakes (1986), Keane and Wolpin (1997), Eckstein and Wolpin (1999), and Hotz and Miller (1988, 1993). Magnac and Thesmar (2002) also consider the case where one state (choice) is absorbing (as do Hotz and Miller (1993)) and where the value functions are known at the terminal age ($\bar{A}$) (as do Keane and Wolpin (1997) and Belzil and Hansen (2002)). In our paper, each treatment time is an absorbing state. In a separate analysis, Magnac and Thesmar consider the case where unobservables from the point of view of the econometrician are correlated over time (or age a) and choices (t) under the assumption that the distribution of the unobservables is known. They also consider the case where exclusion restrictions are available. Throughout their analysis, they maintain that the distribution of the unobservables is known both by the agent and the econometrician.

Our analysis provides semiparametric identification of a finite-horizon finite-state model with an absorbing state with semiparametric specifications of reward and cost functions.[66] Given that rewards are in value units, our utility function cannot be subjected to arbitrary affine transformations so that one source of nonidentifiability in Rust's analysis is eliminated. We can identify the error distributions nonparametrically given our factor structure. We do not have to assume either the functional form of the unobservables or knowledge of the entire distribution of unobservables.

We present a fully specified structural model of choices and outcomes motivated by, but not identical to, the analyses of Keane and Wolpin (1994, 1997) and Eckstein and Wolpin (1999). In their setups, outcome and cost functions are parametrically specified. Their states are recurrent while ours are absorbing. In our model, once an agent drops out of school, the agent does not return. In their model,an agent who drops out can return. They do not establish identification of their model whereas we establish semiparametric identification of our model. We analyze models with more general times series processes for unobservables. In our framework and theirs, agents learn about unobservables. In their framework, such learning is about temporally independent shocks that do not affect agent expectations about

[66] Although our main theorems are for additively separable reward and cost functions, additive separability can be relaxed using the analysis of Matzkin (2003).

returns relevant to possible future choices. The information just affects the opportunity costs of current choices. In our framework, learning affects agent expectations about future returns as well as opportunity costs.

Our model extends previous work by CHH and Cunha, Heckman, and Navarro (2005a,b,c,e) by considering explicit multiperiod dynamic models with information updating. They consider one-shot decision models with information updating and associated outcomes.

Our analysis is related to that of Taber (2000). Like Cameron and Heckman (1998), both our study and Taber's use identification-in-the-limit arguments.[67] Taber considers identification of a two period model with a general utility function whereas in Section 3 we consider identification of a specific form of the utility function (an earnings function) for a multiperiod maximization problem. As in this paper, Taber allows for the sequential arrival of information. His analysis is based on conventional exclusion restrictions, but we do not, as demonstrated in appendix Theorem D.1, text Corollary 1 and in extensions of these results in Section 3. We use outcome data in conjunction with the discrete dynamic choice data to exploit cross equation restrictions, whereas he does not.

Our treatment of unobservables is more general than any discussion that appears in the current dynamic discrete choice and dynamic treatment effect literature. We do not invoke the strong sequential conditional independence assumptions used in the dynamic treatment effect literature in statistics (Robins, 1989, 1997; Gill and Robins, 2001; Lok, 2001; Lechner and Miquel, 2002), nor the closely related conditional temporal independence of unobserved state variables given observed state variables invoked by Rust (1987), Hotz and Miller (1988, 1993), Manski (1993) and Magnac and Thesmar (2002) (in the first part of their paper) or the independence assumptions invoked by Wolpin (1984).[68] We allow for more general

[67] Pakes and Simpson (1989) sketch a proof of identification of a model of the option values of patents that is based on limit sets for an option model.

[68] Manski (1993) and Hotz and Miller (1993) use a synthetic cohort effect approach that assumes that young agents will follow the transitions of contemporaneous older agents in making their lifecycle decisions. The synthetic cohort approach has been widely used in labor economics at least since Mincer (1974). Manski and Hotz and Miller exclude any temporally dependent unobservables from their models. See MaCurdy (1981) and Mincer (1974) for application of the synthetic cohort approach. For empirical evidence against the assumption that the earnings of older workers are a reliable guide to the earnings of younger workers

time series dependence in the unobservables than is entertained by Pakes (1986), Keane and Wolpin (1997) or Eckstein and Wolpin (1999).[69]

Like Miller (1984) and Pakes (1986), we explicitly model, identify and estimate agent learning that affects expected future returns.[70] Pakes and Miller assume functional forms for the distributions of the error process and for the serial correlation pattern about information updating and time series dependence. Our analysis of the unobservables is nonparametric and we estimate, rather than impose, the stochastic structure of the information updating process.

Virtually all papers in the literature, including our own, invoke rational expectations. An exception is the analysis of Manski (1993) who replaces rational expectations with a synthetic cohort assumption that choices and outcomes of one group can be observed (and acted on) by a younger group. This assumption is more plausible in stationary environments and excludes any temporal dependence in unobservables.[71] In recent work, Manski (2004) advocates use of elicited expectations as an alternative to the synthetic cohort approach.

While we use rational expectations, we estimate, rather than impose the structure of agent information sets. Miller (1984), Pakes (1986), Keane and Wolpin (1997), and Eckstein and Wolpin (1999) assume that they know the law governing the evolution of agent information sets up to unknown parameters.[72] Following the procedure presented in Cunha, Heckman, and Navarro (2005a,b,c,e) and Navarro (2004b) we can test for what factors (θ) appear in agent information sets at different stages of the life cycle and we identify the distributions of the unobservables nonparametrically.

Our analysis of dynamic treatment effects is comparable, in some aspects, to the re-

in models of earnings and schooling choices for recent cohorts of workers, see Heckman, Lochner, and Todd (2006).

[69] Rust (1994) provides a clear statement of the stochastic assumptions underlying the dynamic discrete choice literature up to the date of his survey.

[70] As previously noted, the previous literature assumes learning only about current costs.

[71] See Heckman, Lochner, and Todd (2006) for evidence against stationarity assumptions in the analysis of schooling choices for recent cohorts.

[72] They specify *a priori* particular processes of information arrival as well as which components of the unobservables agents know and act on, and which components they do not.

cent continuous time duration analysis of Abbring and Van Den Berg (2003) discussed in Section 3.3. They build a continuous time model of counterfactuals for outcomes that are durations. They model treatment assignment time using a continuous time duration model.

Our analysis is in discrete time and builds on previous work by Heckman (1981a,c) on heterogeneity and state dependence that identifies the causal effect of employment (or unemployment) on future employment (or unemployment).[73] We model time to treatment and associated vectors of outcome equations that may be discrete, continuous or mixed discrete-continuous. In a discrete time setting, we are able to generate a variety of distributions of counterfactuals and economically motivated parameters. We allow for heterogeneity in responses to treatment that has a general time series structure.

As noted in Section 3.3, Abbring and Van Den Berg (2003) do not identify explicit agent information sets as we do in this paper and in Cunha, Heckman, and Navarro (2005e) and they do not model learning about future rewards. Their outcomes are restricted to be continuous time durations. Our discrete time framework avoids many of the technical measure theoretic problems that they and Gill and Robins (2001) encounter in continuous time by using discrete time analysis. We can attach a vector of treatment outcomes that includes continuous outcomes, discrete outcomes and durations expressed as binary strings.[74] At a practical level, we can produce very fine-grained descriptions of continuous time phenomena by using models with many finite periods. Clearly a synthesis of the Abbring – Van Den Berg approach with our approach would be highly desirable. That would entail taking continuous time limits of the discrete time models developed in this paper. It is a task we leave for another occasion.

Flinn and Heckman (1982) utilize information on stopping times and associated wages to use cross equation restrictions to partially identify an equilibrium job search model for a stationary economic environment where agents have an infinite horizon. They establish

[73] Heckman and Borjas (1980) investigate these issues in a continuous time duration model. See also Heckman and MaCurdy (1980).

[74] Abbring (2000) considers nonparametric identification of semi-Markov event history models that extends his work with Van Den Berg.

that the model is nonparametrically nonidentified. Their analysis shows that use of outcome data in conjunction with data on stopping times is not sufficient to secure nonparametric identification. Allowing for nonstationarity arising from finite horizons can break their nonidentification result (see Wolpin, 1987). Our analysis exploits the finite-horizon backward-induction structure of our model in conjunction with outcome data to secure identification and does not rely on arbitrary period by period exclusion restrictions. We substantially depart from the assumptions maintained in Rust's nonidentification theorem (1994). We achieve identification by using more information and exploiting the structure of our finite horizon nonrecurrent model. Nonstationarity of regressors greatly facilitates identification by producing both exclusion and curvature restrictions which can substitute for exclusion restrictions. We leave exploration of identification of an infinite horizon version of our model with recurrent states in a stationary environment for another occasion.

5 Conclusion

This paper develops two econometric models of time to treatment (or dropout) and associated systems of outcomes generated at different treatment times. A third benchmark model for a conventional static discrete choice framework with counterfactuals is developed in Appendix B. Our semiparametric analysis of a dynamic discrete choice model with associated outcomes allows for general time series processes for the unobservables and agent learning. We do not make parametric assumptions about model unobservables. The outcomes we analyze may be discrete, continuous or mixed discrete-continuous random variables, although in this paper we focus on the continuous outcome case in analyzing structural models. We establish conditions for semiparametric identification of these models, and we develop the counterfactuals that can be produced by each model. Our identification analysis of the time to treatment is of interest in its own right and constitutes an independent contribution to the semiparametric analysis of dynamic discrete choice models. Our explicit choice

theoretic model is suitable for the analysis of outcomes associated with different times to treatment in conjunction with choice data on times to treatment. The cross-equation restrictions generated by choice theory and the nonstationarity induced by agent finite horizons help to identify agent preferences (costs) and agent information sets. Access to measurement equations is helpful in identifying the unobservables associated with cost functions, and in constructing distributions of outcomes across stopping times, measurements are not needed for identification of choice equations or of state-specific outcome equations. We identify *ex ante* and *ex post* objective and subjective evaluations of outcomes and allow for updating of expected rewards and stopping times as information accumulates over the life cycle.

The reduced form models we analyze cannot identify treatment effects motivated by choice theory such as the marginal treatment effect (MTE). They also generate certain counterfactuals that are difficult to interpret and can violate basic principles of causality. The benchmark multinomial discrete choice model with associated outcomes developed in Appendix B rules out option values but that can produce all of the conventional *ex post* treatment effects.

Heckman and Navarro (2006) present estimates of option values and compare the predictive performance of static and structural models. Cunha, Heckman, and Navarro (2005d) consider identification of a generalized ordered discrete choice model with stochastic thresholds that rules out many of the perversities associated with the unrestricted reduced form time to treatment model but at the cost of eliminating option values. Our paper demonstrates the value of articulated economic choice models in elucidating the structure of statistical treatment effect models and in identifying parameters of costs, preferences and returns.

Appendices

A The Matzkin Conditions

Consider a binary choice model, $D = \mathbf{1}(\varphi(Z) > V)$, where Z is observed and V is unobserved. Let φ^* denote the true φ and let F_V^* denote the the true cdf of V. Let $Z \in \mathcal{Z}$. Let Γ denote the set of monotone increasing functions from $\mathbb{R}$ into $[0, 1]$. Assume

(i) $\varphi^* : \mathcal{Z} \to \mathbb{R}$, where $\mathcal{Z} \subset \mathbb{R}^K$ and $\varphi^* \in \Phi$, where Φ is a set of functions mapping $\mathcal{Z}$ into $\mathbb{R}$ that are continuous and strictly increasing in their K^{th} coordinate.

(ii) $Z \perp\!\!\!\perp V$

(iii) The conditional distribution of the K^{th} coordinate of Z has a Lebesgue density that is everywhere positive conditional on the other coordinates of Z.

(iv) F_V^* is strictly increasing.

(v) The support of the marginal distribution of Z is included in $\mathcal{Z}$.

Then (φ^*, F_V^*) is identified within $\Phi \times \Gamma$ if and only if Φ is a set of functions such that no two functions in Φ are strictly increasing transformations of each other (Matzkin, 1994).

She also shows that the following alternative representations of functional forms satisfy the conditions for exact identification for $\varphi(Z)$.

1. $\varphi(Z) = Z\gamma$, $\|\gamma\| = 1$ or $\gamma_1 = 1$.

2. $\varphi(Z)$ is homogeneous of degree one attains a given value α, at $Z = z^*$ (*e.g.* cost functions).

3. Least concave functions that attain common values at two points in their domain.

4. Additively separable functions:

(a) Functions additively separable into a continuous and monotone increasing function and a continuous monotone increasing, concave and homogeneous of degree one function.

(b) Functions additively separable into the value of one variable and a continuous, monotone increasing function of the remaining variables

(c) Additively separable functions, *e.g.* $\varphi(Z) = Z_1 + \tau(Z_2, \ldots, Z_K)$

B Identification of Counterfactual Outcomes for a Multinomial Discrete Choice Model with State-Contingent Outcomes

Let outcomes in state s be $Y(s, X) = \mu_Y(s, X) + U(s)$, $s = 1, \ldots, \bar{S}$, where there are $\bar{S}$ discrete states. Let $V(s, Z) = \mu_V(s, Z) + \eta(s)$. The $U(s)$ and $\eta(s), s = 1, \ldots, \bar{S}$ are assumed to be continuous and measurably separated as a collection of random variables. Thus the support of one random variable does not restrict the supports of the other random variables. State s is selected if

$$s = \underset{j}{\operatorname{argmax}} \{V(j, Z)\}_{j=1}^{\bar{S}}$$

and $Y(s, X)$ is observed. If s is observed, $D(s) = 1$. Otherwise $D(s) = 0$. $\sum_{s=1}^{\bar{S}} D(s) = 1$. Matzkin (1993) considers identification of polychotomous discrete choice models under the conditions of the Theorem B.1 below. We extend her analysis by adjoining counterfactual outcomes associated with each choice. We can identify $\mu_Y(s, X), s = 1, \ldots, \bar{S}$ over the support of X; $\mu_V(s, Z)$, up to scale over the support of Z and the joint distributions of $\big(U(s), \eta(s) - \eta(1), \ldots, \eta(s) - \eta(s-1), \eta(s) - \eta(s+1), \ldots, \eta(s) - \eta(\bar{S})\big)$ with the contrasts $\eta(s) - \eta(\ell)$ up to a scale that we present below in our discussion of Theorem B.1.

Theorem B.1. *Assume*

(i) $\left(U\left(1\right),\ldots,U\left(\bar{S}\right),\eta(1),\ldots,\eta(\bar{S})\right)$ *are continuous random variables (absolutely continuous with respect to Lebesgue measure).*

(ii) They are measurably separated random variables so that $Supp\left(U\left(s\right),\eta(1),\ldots,\eta(\bar{S})\right) = Supp\left(U\left(s\right)\right) \times Supp\left(\eta(1)\right) \times \cdots \times Supp\left(\eta(\bar{S})\right)$

(iii) $Supp\left(\mu_Y\left(s,X\right),\mu_V\left(s,X\right)-\mu_V\left(1,X\right),\ldots,\mu_V\left(s,X\right)-\mu_V\left(\bar{S},X\right)\right) = Supp\left(\mu_Y\left(s,X\right)\right) \times \prod_{\substack{s'=1 \\ s'\neq s}}^{\bar{S}} Supp\left(\mu_V\left(s,X\right)-\mu_V\left(s',X\right)\right), \quad s=1,\ldots,\bar{S}$

(iv) $Supp\left(\mu_Y\left(s,X\right),\mu_V\left(s,X\right)-\mu_V\left(1,X\right),\ldots,\mu_V\left(s,X\right)-\mu_V\left(\bar{S},X\right)\right) \supseteq Supp\left(U\left(s\right),\eta(s)-\eta(1),\ldots,\eta(s)-\eta(\bar{S})\right),\ s=1,\ldots,\bar{S}$

(v) $\left(U\left(s\right),\eta(s)-\eta(1),\ldots,\eta(s)-\eta(\bar{S})\right) \perp\!\!\!\perp (X,Z)\ s=1,\ldots,\bar{S}$

Then $\mu_Y\left(s,X\right), s=1,\ldots,\bar{S}$, *is identified;* $\left(\mu_V\left(s,X\right)-\mu_V\left(1,X\right),\ldots,\mu_V\left(s,X\right)-\mu_V\left(\bar{S},X\right)\right)$, *are identified up to a common scale for all* $s=1,\ldots,\bar{S}$, *and the distribution of* $(U\left(s\right),\eta(s)-\eta(1),\ldots,\eta(s)-\eta(\bar{S}))$ *is identified, the last* $\bar{S}-1$ *components up to a common scale.*

Proof. This theorem follows from an application of Theorem 3 in CHH. Because of (iii) we can find limit sets $\mathcal{Z}$ such that

$$\lim_{Z\to\mathcal{Z}} \Pr\left(D\left(s\right)=1 \mid Z\right)=1$$

and we can identify the $\mu_Y\left(s,X\right), s=1,\ldots,\bar{S}$ in those limit sets. We can then vary $\mu_Y\left(s,X\right)$ and trace out the marginal distribution of the $U\left(s\right), s=1,\ldots,\bar{S}$. By similar reasoning, we identify the $\left(\mu_V\left(s,X\right)-\mu_V\left(1,X\right),\ldots,\mu_V\left(s,X\right)-\mu_V\left(\bar{S},X\right)\right)$ up to scale. We can, by virtue of (iv), trace out the joint distribution of $\left(U\left(s\right),\eta(s)-\eta(1),\ldots,\eta(s)-\eta(\bar{S})\right)$, $s=1,\ldots,\bar{S}$ with the last $\bar{S}$ coordinates identified up to scale on the unobservables. ■

Invoking the Matzkin conditions we can set the scale of the deterministic functions. If we invoke her functions up to an unknown scale, we only identify the functions up to scale. We identify the $\mu_Y\left(s,X\right)$ and the scaled version of

$(\mu_V(s,X)-\mu_V(1,X),\ldots,\mu_V(s,X)-\mu_V(\bar{S},X))$ over the supports of X and Z respectively. Exclusion restrictions are the traditional way to satisfy conditions (iii) and (iv) but these are not required as the argument of Corollary 1 of Theorem 1 proved in Appendix C demonstrates. With minor modification, the proof structure of this corollary can be adapted to this setting. Matzkin (1993) provides conditions for identification of the $V(j,Z)$ in the random utility case with conventional structure.

From this model, we can identify the marginal treatment effect (CHH, p. 368, equation (71)) and all pairwise average treatment effects by forming suitable limit sets. We can also identify all pairwise mean treatment on the treated and mean treatment on the untreated effects.

In the general case, we can identify the densities of $U(s),\eta(s)-\eta(1),\ldots,\eta(s)-\eta(\bar{S})$, $s=1,\ldots,\bar{S}$,where $U(s)$ may be a vector and the contrasts are identified up to a scale which we now define. Set $Var(\eta(s))=1$ for all $s=1,\ldots,\bar{S}-1$. Set $\mu_V(\bar{S},Z)\equiv 0$ and $\eta(\bar{S})\equiv 0$.[75] From the choice equation for $\bar{S}$ $\left(\Pr\left(D(\bar{S})=1\mid Z=z\right)\right)$, we can identify the pairwise correlations $\rho_{i,j}=Correl(\eta(i),\eta(j))$, $i,j=1,\ldots,\bar{S}-1$. We assume that $-1\leq\rho_{i,j}<1$. If $\rho_{i,j}=1$ for some i,j, the choice of a normalization is not innocuous. Under our assumptions, we can identify $Var(\eta(s)-\eta(\ell))=2\left(1-\rho_{s,\ell}\right)$. Define $\tau_{s,\ell}=[Var(\eta(s)-\eta(\ell))]^{1/2}$ where positive square roots are used. This is used to set the scale for contrast s,ℓ.

Consider constructing the distribution of $Y(\ell,X)$ given $D(s)=1,X,Z$. If $\ell\neq s$, this is a counterfactual distribution. From this distribution we can construct, among many possible counterfactual parameters, $E(Y(s,X)-Y(\ell,X)\mid D(s)=1,X=x,Z=z)$, a treatment on the treated parameter. We can also construct

$$E\left(Y(s,X)-Y(\ell,X)\;\middle|\;\begin{array}{l}V(s,Z)=V(\ell,Z),\\ V(s,Z),V(\ell,Z)\geq\max\limits_{\substack{j=1,\ldots,\bar{S}\\ j\neq s,\ell}}\{V(j,Z)\}\end{array}\right),$$

the effect of moving from state ℓ to state s for people at the margin of indifference between

[75] This is one of many possible normalizations.

s and ℓ.[76]

To form the counterfactual distribution $\left(U(\ell), \frac{(\eta(s)-\eta(1))}{\tau_{s,1}}, \ldots, \frac{\left(\eta(s)-\eta(\bar{S})\right)}{\tau_{s,\bar{S}}}\right)$ for any $\ell \neq s$ from the output of Theorem B.1, we use the normalized versions of $\eta(s)-\eta(1), \ldots, \eta(s)-\eta(\bar{S})$: $\frac{(\eta(s)-\eta(1))}{\tau_{s,1}}, \ldots, \frac{\left(\eta(s)-\eta(\bar{S})\right)}{\tau_{s,\bar{S}}}$. From the density of $U(\ell), \frac{(\eta(\ell)-\eta(1))}{\tau_{\ell,1}}, \ldots, \frac{\left(\eta(\ell)-\eta(\bar{S})\right)}{\tau_{\ell,\bar{S}}}$ which we identify from Theorem B.1, we can transform the contrast variables in the following way.

Define $q(\ell, s) = \frac{(\eta(\ell)-\eta(s))}{\tau_{\ell,s}}$. Observe that $q(s,j) = \frac{\eta(s)-\eta(j)}{\tau_{s,j}} = \frac{q(\ell,j)\tau_{\ell,j}-q(\ell,s)\tau_{\ell,s}}{\tau_{s,j}}$ for all $j = 1, 2, \ldots, \bar{S}$. Replace $\frac{\eta(s)-\eta(j)}{\tau_{s,j}}$ by $\frac{q(\ell,j)\tau_{\ell,j}-q(\ell,s)\tau_{\ell,s}}{\tau_{s,j}}$ $j = 1, 2, \ldots, \bar{S}$, $j \neq \ell$ in the density of $(U(\ell), \frac{(\eta(\ell)-\eta(s))}{\tau_{\ell,s}}, \ldots, \frac{\left(\eta(\ell)-\eta(\bar{S})\right)}{\tau_{\ell,\bar{S}}})$ and use the Jacobian of transformation $\prod_{j=1,\ldots,\bar{S},j\neq\ell} |\tau_{\ell,j}|$, where "$|\quad|$" denotes determinant. Thus we can generate the desired counterfactual density for all $s = 1, \ldots, \bar{S}$. Provided that the Jacobians are nonzero (which rules out perfect dependence, $\rho_{\ell,s} \neq 1$, $\ell \neq s$), we preserve all of the information and can construct the marginal distribution of any $U(\ell)$ for any desired pattern of latent indices. Thus we can construct the desired counterfactuals.

The key difference between this model and the one developed in Section 2 in the text is that across all counterfactual states the same collection of random variables generates the $D(s)$, $s = 1, \ldots, \bar{S}$. In contrast, in the model of Sections 2 and Section 3, new random variables are added at each stage of the time to treatment process. If we control the proliferation of unobservables, as we do in the factor model of Section 2.5, we can identify all of the traditional counterfactual means and the distributions of outcomes as well.

[76] Heckman (2006) and Heckman and Vytlacil (2006a) call this parameter *EOTM*, the effect of treatment for people at the margin.

C Identification Proofs

Proof. (Theorem 1) Let

$$\begin{aligned} S_{\eta(1)}(z(1)\gamma_1) &= 1 - F_{\eta(1)}(z(1)\gamma_1) \\ &= 1 - \Pr(D(1) = 1 \mid Z(1) = z(1)) \\ &= \Pr(D(1) = 0 \mid Z(1) = z(1)) \\ &= \Pr\left(z(1)\gamma_1 < \eta\left(1\right) \mid Z(1) = z(1)\right). \end{aligned}$$

Similarly let

$$S_{\eta(1),\eta(2)}(z(1)\gamma_1, z(2)\gamma_2) = \Pr\left(z(1)\gamma_1 < \eta\left(1\right) \wedge z(2)\gamma_2 < \eta\left(2\right) \mid Z(1) = z(1), Z(2) = z(2)\right)$$

and so forth. By hypothesis, we know the left hand sides of the following $\bar{T}$ equations:

$$\begin{aligned} \Pr\left(D\left(1\right) = 0 \mid Z(1) = z(1)\right) &= S_{\eta(1)}\left(z(1)\gamma_1\right) \qquad \text{(C.1)} \\ \Pr\left(D\left(1\right) = 0, D\left(2\right) = 0 \mid Z(1) = z(1), Z(2) = z(2)\right) &= S_{\eta(1),\eta(2)}\left(z(1)\gamma_1, z(2)\gamma_2\right) \\ &\vdots \\ \Pr\left(\begin{array}{c|c} D\left(1\right) = 0, D(2) = 0, \ldots, & Z(1) = z(1), \ldots, \\ D\left(\bar{T}\right) = 0 & Z(\bar{T}) = z(\bar{T}) \end{array}\right) &= S_{\eta(1),\ldots,\eta\left(\bar{T}\right)}\left(z(1)\gamma_1, \ldots, z(\bar{T})\gamma_{\bar{T}}\right). \end{aligned}$$

We may treat the first equation as a binary discrete choice model. Following the analysis of Manski (1988, Proposition 2, Corollary 5), under the conditions of the theorem we can identify γ_1 and $S_{\eta(1)}$ up to scale and location. For example, we may normalize the location and scale by assuming $E(\eta\left(1\right)) = 0$ and by requiring that $\|\gamma_1\| = 1$, where $\|\gamma_1\|$ is the norm of the vector γ_1.

We cannot directly apply Manski's analysis for $T \geq 2$. We do not directly observe $\Pr(D\left(2\right) = 0 \mid Z(2))$, since the $D\left(2\right)$ outcome is not observed for individuals with $D\left(1\right) = 1$.

We therefore proceed with a recursive "identification in the limit" argument.

If the true parameter values are $\left(S_{\eta^0(2)}, \gamma_2^0\right)$, then given the identification of the first period parameters which we just established, the second period parameters are identified, iff for any alternative parameter values $\left(S_{\eta^*(2)}, \gamma_2^*\right) \in \Gamma_2 \times \mathcal{H}_2$ with $(S_{\eta^*(2)}, \gamma_2^*) \neq (S_{\eta^0(2)}, \gamma_2^0)$, there exists some $\varphi > 0$ such that

$$\Pr_{Z|D(1)=0}\left(\left|S_{\eta^0(1),\eta^0(2)}\left(Z(1)\gamma_1^0, Z(2)\gamma_2^0\right) - S_{\eta^0(1),\eta^*(2)}\left(Z(1)\gamma_1^0, Z(2)\gamma_2^*\right)\right| > \varphi\right) > 0. \quad \text{(C.2)}$$

Pick any $\left(S_{\eta^*(2)}, \gamma_2^*\right) \in \Gamma_2 \times \mathcal{H}_2 \setminus \left(S_{\eta^0(2)}, \gamma_2^0\right)$. We now show that (C.2) holds for some $\varphi > 0$.

By continuity of $S_{\eta^0(1)}$, for any $\varepsilon > 0$ we can pick $\tilde{g}_1 \in \left(\underline{\eta}(1), \bar{\eta}(1)\right)$ such that

$$S_{\eta^0(1)}(g_1) \leq \varepsilon/2 \text{ for all } g_1 \geq \tilde{g}_1 \Longrightarrow \sup_{g_2}\left|S_{\eta^0(1),\eta^0(2)}(g_1, g_2) - S_{\eta^0(2)}(g_2)\right| \leq \varepsilon/2 \quad \text{(C.3)}$$

and

$$\sup_{g_2}\left|S_{\eta^0(1),\eta^*(2)}(g_1, g_2) - S_{\eta^*(2)}(g_2)\right| \leq \varepsilon/2 \quad \text{(C.4)}$$

for all $g_1 \geq \tilde{g}_1$. The triangle inequality implies that

$$\left|\begin{array}{r} \left[S_{\eta^0(1),\eta^0(2)}\left(Z(1)\gamma_1^0, Z\left(2\right)\gamma_2^0\right) - S_{\eta^0(1),\eta^*(2)}\left(Z(1)\gamma_1^0, Z\left(2\right)\gamma_2^*\right)\right] \\ \\ - \left[S_{\eta^0(2)}\left(Z(2)\gamma_2^0\right) - S_{\eta^*(2)}\left(Z(2)\gamma_2^*\right)\right] \end{array}\right| \leq \varepsilon. \quad \text{(C.5)}$$

From this, it follows that

$$\Pr\left(\left|\begin{array}{c} S_{\eta^0(1),\eta^0(2)}\left(Z(1)\gamma_1^0, Z(2)\gamma_2^0\right) \\ -S_{\eta^0(1),\eta^*(2)}\left(Z(1)\gamma_1^0, Z(2)\gamma_2^*\right) \end{array}\right| > \varphi \;\middle|\; Z(1)\gamma_1^0 \geq \max(\tilde{g}_1, \breve{g}_1)\right)$$

$$\geq \Pr\left(\left|S_{\eta^0(2)}\left(Z(2)\gamma_2^0\right) - S_{\eta^*(2)}\left(Z(2)\gamma_2^*\right)\right| > \varphi + \varepsilon \;\middle|\; Z(1)\gamma_1^0 \geq \max(\tilde{g}_1, \breve{g}_1)\right).^{77}$$

Using conditions (iii) and (iv) of the Theorem, $\Pr(S_{\eta^0(2)}(Z(2)\gamma_2^0) = S_{\eta^*(2)}(Z(2)\gamma_2^*) \mid Z(1)\gamma_1^0 \geq \max(\tilde{g}_1, \check{g}_1)) = 1$ iff $(S_{\eta^*(2)}, \gamma_2^*) = (S_{\eta^0(2)}, \gamma_2^0)$. Since $(S_{\eta^*(2)}, \gamma_2^*) \neq (S_{\eta^0(2)}, \gamma_2^0)$, and since we can set ε arbitrarily small, there exists φ values such that the last probability is strictly positive so that, for such φ values,

$$\Pr\left(\left|\begin{array}{c} S_{\eta^0(1),\eta^0(2)}\left(Z(1)\gamma_1^0, Z\left(2\right)\gamma_2^0\right) \\ -S_{\eta^0(1),\eta^*(2)}\left(Z(1)\gamma_1^0, Z\left(2\right)\gamma_2^*\right) \end{array}\right| > \varphi \;\middle|\; Z(1)\gamma_1^0 \geq \max(\tilde{g}_1, \check{g}_1)\right) > 0.$$

Using (iv), we have that the conditioning set in (C.2) has positive probability

$$\Pr(Z(1)\gamma_1^0 \geq \max(\tilde{g}_1, \check{g}_1)) > 0,$$

so that (C.2) holds. We have shown that $(S_{\eta^*(2)}, \gamma_2^*) \neq (S_{\eta^0(2)}, \gamma_2^0)$ implies (C.2), and thus the $(S_{\eta^0(2)}, \gamma_2^0)$ parameters are identified. Proceeding in this fashion, we can recover $Z(t)\gamma_t^0$, $t = 1, \ldots, \bar{T}$. Since we identify $Z(t)\gamma_t^0$ using (iv), we can recover the joint distribution of $(\eta(1), \ldots, \eta(\bar{T}))$ varying the components of $(Z(1)\gamma_1^0, \ldots, Z(\bar{T})\gamma_{\bar{T}}^0)$ to trace out $S_{\eta(1),\ldots,\eta(\bar{T})}$ and hence we can recover $F_{\eta(1),\ldots,\eta(\bar{T})}$. ■

Proof. (Corollary 1) Let

$$z\gamma_1 = g_1.$$

Recall that $\|\gamma_t\| = 1$ for some $t = 1, \ldots, T^*$, is our normalization. The first T^* coordinates of z correspond to continuous regressors. By assumption (vi), $\gamma_{11} \neq 0$, and we can write

$$z_1 = \frac{g_1}{\gamma_{11}} - z_2\frac{\gamma_{12}}{\gamma_{11}} - \cdots - z_K\frac{\gamma_{1K}}{\gamma_{11}}$$

where in this expression, lower case z_i is the i^{th} coordinate of z.

In the index $z\gamma_2$ use Gaussian elimination and substitute for z_1 from the preceding

[77] The intuition for this result is that if the first term inside (C.5) is bigger than φ in absolute value, the second term in (C.5) must be within $\varphi \pm \varepsilon$ in absolute value since the two terms live in a narrow band defined by ε.

equation to obtain the expression

$$\left(\frac{g_1}{\gamma_{11}} - z_2\frac{\gamma_{12}}{\gamma_{11}} - \cdots - z_K\frac{\gamma_{1K}}{\gamma_{11}}\right)\gamma_{21} + \gamma_{22}z_2 + \cdots + \gamma_{2K}z_K. \tag{C.6}$$

Under assumption (*iii*) of Theorem 1 as amended in assumption (*v*) in the statement of Corollary 1, these variables can be freely varied given $z\gamma_1 = g_1$. Proceeding recursively, in the $(j+1)^{th}$ argument, $(j < T^*)$, we obtain an expression that substitutes out for $(z_1, \ldots, z_j)$ leaving $T^* - j$ free continuous variables and $\bar{T} - j$ total remaining variables.

Array the γ_j into a matrix C with the j^{th} column of C being γ_j. C is a $K \times \bar{T}$ matrix. Let $C(r, n)$ be the $n \times r$ submatrix of C consisting of the first n rows and r columns, and let $C(r, K-n)$ be the matrix consisting of the last $K-n$ rows and the first r columns of C. Partition γ_j into the first e elements $\left(\gamma_j(e)\right)$ and the last $K-e$ elements $\gamma_j(K-e)$. Finally, let $\tilde{z}_j$ be the last $\bar{T} - j$ elements of z and $\tilde{\gamma}_j$ denote the parameters associated with them at the j^{th} step of the Gaussian elimination process.

In this notation, the index $z\gamma_2$ in equation (C.6) can be written as

$$g_1\frac{\gamma_{21}}{\gamma_{11}} + \tilde{z}_2\tilde{\gamma}_2.$$

Successive Gaussian elimination produces

$$\tilde{\gamma}_{j+1} = \gamma_{j+1}(K-j) - C(j, K-j)\left[C(j,j)\right]^{-1}\gamma_{j+1}(j)$$

a $K-j$ dimensional vector. In order for $[C(j,j)]^{-1}$ to exist, $j = 1, \ldots, T^*$, it is necessary that $\gamma_1, \ldots, \gamma_j$ be linearly independent vectors. Condition (*v*) assures us that this requirement is satisfied for $j \leq T^*$. Define $\tilde{\gamma}_{j+1}(T^* - j)$ as the first $(T^* - j)$ elements of $\tilde{\gamma}_{j+1}$ associated with the continuous regressors. In order to satisfy (*vi*), at least one component of $\tilde{\gamma}_{j+1}(T^* - j)$ must be non-zero.

Again consider

$$g_1 = z\gamma_1$$

$$g_2 = \tilde{\gamma}_2(g_1) + \tilde{z}_2\tilde{\gamma}_2$$

where $\tilde{\gamma}_2(g_1) = \left(\frac{g_1}{\gamma_{11}}\gamma_{21}\right)$ is obtained using the same linear transformation that is used to obtain $\tilde{\gamma}_{j+1}$ with $j = 1$. Since $\tilde{\gamma}_2(g_1)$ is a function of g_1, the second period index is a function of g_1 and for fixed $\tilde{z}_2$ we have that $g_1 \to \infty \Longrightarrow g_2 \to \infty$. However, note that using assumptions *(iii)* and *(v)* of Theorem 1 and assumptions (v) and (vi) of the corollary, we can send $g_1 \to \infty$ while varying $\tilde{z}_2$ to keep g_2 fixed. In particular, we can use z_1 to send $g_1 \to \infty$ and set z_2 to compensate for z_1 in the second period index so as to hold g_2 fixed. Thus, $Supp(Z\gamma_2|Z\gamma_1 = g_1) = R$ and the Z that satisfy $Z\gamma_1 = g_1$ will have rank $K-1$ for *a.e.* $g_1 \in R$. Moreover, we have, for *a.e.* $g_1 \in R$, $Supp(Z\gamma_2|Z\gamma_1 \geq g_1) = R$ and the Z that satisfy $Z\gamma_1 \geq g_1$ has full rank (there exists no proper linear subspace of R^K having probability 1 under $F_{Z|Z\gamma_1 \geq g_1}$). We can repeat this argument, using sequential Gaussian elimination as described above, to show that

$$Supp\left(Z\gamma_t|Z\gamma_1 = g_1, \ldots, Z\gamma_{t-1} = g_{t-1}\right) = R, \qquad t \leq T^*,$$

and there exists no proper linear subspace of R^K having probability 1 under $F_{Z|Z\gamma_1 \geq g_1, \ldots, Z\gamma_t \geq g_t}$ for almost every $(g_{t-1}, \ldots, g_1) \in R^{t-1}$ for $t = 2, \ldots, \bar{T}$. Using the argument in Cameron and Heckman (1998), we can identify all the remaining parameters of the model (γ_t, for $t = 1, \ldots, T^*$, up to scale and location normalizations). ■

D Identification of the General Model of Section 2

This appendix generalizes the analysis of Theorems 2 and 3 in the text. Use $Y(a,t)$ as shorthand for $Y(a,t,X,U(a,t))$. Ignore (for notational simplicity) the mixed discrete-continuous outcome case. We can build that case from the continuous and discrete cases and for the sake

of brevity we do not analyze it here. We also do not analyze duration outcomes although it is straightforward to do so.[78] We decompose $Y(a,t)$ into discrete and continuous components:

$$Y(a,t) = \begin{bmatrix} Y_c(a,t) \\ Y_d(a,t) \end{bmatrix}.$$

Associated with the j^{th} component of $Y_d(a,t)$, $Y_{d,j}(a,t)$ is a latent variable $Y^*_{d,j}(a,t)$. We define

$$Y_{d,j}(a,t) = \mathbf{1}\left(Y^*_{d,j}(a,t) \geq 0\right).^{79}$$

From standard results in the discrete choice literature, without additional information, we can only know $Y^*_{d,j}(a,t)$ up to scale.

We assume an additively separable model for the continuous variables and latent continuous indices. Making the X explicit, we have

$$\begin{aligned} Y_c(a,t,X) &= \mu_c(a,t,X) + U_c(a,t) \\ Y^*_d(a,t,X) &= \mu_d(a,t,X) - U_d(a,t) \\ 1 &\leq t \leq \bar{T}, 1 \leq a \leq \bar{A}. \end{aligned}$$

We array the $Y_c(a,t,X)$ into a matrix $Y_c(t,X)$ and the $Y^*_d(a,t,X)$ into a matrix $Y^*_d(t,X)$. We decompose these vectors into components corresponding to the means $\mu_c(t,X), \mu_d(t,X)$ and the unobservables $U_c(t), U_d(t)$. Thus

$$\begin{aligned} Y_c(t,X) &= \mu_c(t,X) + U_c(t) \\ Y^*_d(t,X) &= \mu_d(t,X) - U_d(t). \end{aligned}$$

[78] The ingredients for doing so are in Corollary 2 of Theorem 3

[79] Extensions to nonbinary discrete outcomes are straightforward. Thus we could entertain, at greater notational cost, a multinomial outcome model at each age a for each counterfactual state, building on the analysis of Appendix B.

$Y_d^*(t,X)$ generates $Y_d(t,X)$. To simplify the notation, we will make use of the condensed forms $Y_c(X)$, $Y_d^*(X)$, $\mu_c(X)$, $\mu_d(X)$, U_c and U_d as described in Section 2.3. In this notation,

$$\begin{aligned} Y_c(X) &= \mu_c(X) + U_c \\ Y_d^*(X) &= \mu_d(X) - U_d. \end{aligned}$$

Following CHH, and Cunha, Heckman, and Navarro (2005a,b,c,e), we may also have a system of measurements with both discrete and continuous components. The measurements are not t-indexed. They are the same for each stopping time. (Hansen, Heckman, and Mullen, 2004, generalize a version of the model discussed in this section to allow for t-specific measurements.) We write the equations for the measurements in an additively separable form, in a fashion comparable to those of the outcomes. The equations for the continuous measurements and latent indices producing discrete measurements are

$$\begin{aligned} M_c(a,X) &= \mu_{c,M}(a,X) + U_{c,M}(a) \\ M_d^*(a,X) &= \mu_{d,M}(a,X) - U_{d,M}(a) \end{aligned}$$

where the discrete variable corresponding to the j^{th} index in $M_d^*(a,X)$ is

$$M_{d,j}(a,X) = \mathbf{1}\left(M_{d,j}^*(a,X) \geq 0\right).$$

The measurements play the role of indicators unaffected by the process being studied. We array $M_c(a,X)$ and $M_d^*(a,X)$ into matrices $M_c(X)$ and $M_d^*(X)$. We array $\mu_{c,M}(a,X)$, $\mu_{d,M}(a,X)$ into matrices $\mu_{c,M}(X)$ and $\mu_{d,M}(X)$. We array the corresponding unobservables into $U_{c,M}$ and $U_{d,M}$. Thus we write

$$\begin{aligned} M_c(X) &= \mu_{c,M}(X) + U_{c,M} \\ M_d^*(X) &= \mu_{d,M}(X) - U_{d,M}. \end{aligned}$$

We use the notation of Section 2.4 to write $I(t) = \Psi(t, Z) - \eta(t)$ and collect $I(t), \Psi(t, Z)$ and $\eta(t)$ into vectors I, $\mu(Z)$, η. We define $\eta^t = (\eta(1), \ldots, \eta(t))$ and $\Psi^t(Z) = (\Psi(1, Z), \ldots, \Psi(t, Z))$. Using this notation, we extend the analysis of CHH to identify our model assuming that (Y_c, Y_d, M_c, M_d, I) are independently distributed across people.

Theorem D.1. *The joint distribution of* $(U_c(t), U_d(t), U_{c,M}, U_{d,M}, \eta^t)$ *is identified (the components corresponding to discrete outcomes up to scale) along with the mean functions* $(\mu_c(t, X), \mu_d(X), \mu_{c,M}(X), \mu_{d,M}(X), \Psi^t(Z))$ *with mean functions for the* $\Psi^t(Z)$ *and the discrete outcome components belonging to the Matzkin class of functions if*

(i) $(U_c, U_d, U_{c,M}, U_{d,M}, \eta^t)$ *are continuous random variables with zero means, finite variances and support:* $Supp(U_c) \times Supp(U_d) \times Supp(U_{c,M}) \times Supp(U_{d,M}) \times Supp(\eta^t)$ *with upper and lower limits* $\left(\bar{U}_c, \bar{U}_d, \bar{U}_{c,M}, \bar{U}_{d,M}, \bar{\eta}^t\right)$ *and* $\left(\underline{U}_c, \underline{U}_d, \underline{U}_{c,M}, \underline{U}_{d,M}, \underline{\eta}^t\right)$ *respectively. These conditions are assumed to apply within each component of each subvector. The joint system is thus measurably separable (variation free) for each component with respect to every other component.*

(ii) $(U_c, U_d, U_{c,M}, U_{d,M}, \eta^t) \perp\!\!\!\perp (X, Z)$.

(iii) $Supp\left(\mu_c(t, X), \mu_d(t, X), \mu_{c,M}(X), \mu_{d,M}(X), \Psi^t(Z)\right) =$

$Supp(\mu_c(t, X)) \times Supp(\mu_d(t, X)) \times Supp\left(\mu_{c,M}(X)\right) \times Supp\left(\mu_{d,M}(X)\right) \times Supp(\Psi^t(Z))$

and a comparable condition holds for all subcomponents;

(iv) $Supp\left(\mu_d(t, X), \mu_{d,M}(X), \Psi^t(Z)\right) \supseteq Supp(U_d(t), U_{d,M}, \eta^t)$,

where $\eta^t = (\eta(1), \ldots, \eta(t))$ *collects the first* t *elements of* η.

Proof. From the data on $Y_c(t, X), Y_d(t, X), M_c(X), M_d(X)$ for $D(t) = 1, D^{t-1} = (0)$, and from the time to treatment probabilities, we can construct the left hand side of the following

equation:

$$\Pr\left(\begin{array}{l} Y_c(t,X) \leq y_c(t,X), \mu_d(t,X) \leq U_d(t), \\ M_c(X) \leq m_c(X), \mu_{d,M}(t,X) \leq U_{d,M} \end{array} \middle| \; D(t)=1, D^{t-1}=(0), X=x, Z=z \right)$$

$$\times \Pr\left(D(t)=1, D^{t-1}=(0) \mid X=x, Z=z\right)$$

$$= \int_{\underline{U}_c}^{y_c(t,x)-\mu_c(t,x)} \int_{\mu_d(t,x)}^{\bar{U}_d} \int_{\underline{U}_{c,M}}^{m_c(x)-\mu_{c,M}(x)} \int_{\mu_{d,M}(x)}^{\bar{U}_{d,M}} \tag{D.1}$$

$$\int_{\underline{\eta}_t}^{\Psi(t,z)} \int_{\Psi(t-1,z(t-1))}^{\bar{\eta}(t-1)} \cdots \int_{\Psi(1,z(1))}^{\bar{\eta}_1} f_{U_c(t),U_d(t),U_{c,M},U_{d,M},\eta^t}\left(u_c(t), u_d(t), u_{c,M}, u_{d,M}, \eta(1), \ldots, \eta(t)\right)$$
$$\cdot d\eta(1) \cdots d\eta(t)\, du_{d,M}\, du_{c,M}\, du_d(t)\, du_c(t)$$

(Recall that $D(0)=0$ is fixed outside the model.)

Under assumptions *(i)-(iv)*, for all $x \in Supp(X)$, we can vary the $\Psi(j,Z)$, $j=1,\ldots,t$ and obtain a limit set $\mathcal{Z}$ such that $\lim_{z \to \mathcal{Z}} \Pr\left(D(t)=1, D^{t-1}=(0) \mid X=x, Z=z\right)=1$. We can identify the joint distribution of $Y_c(t,X), Y_d(t,X), M_c(X), M_d(X)$ free of selection bias for all $t=1,\ldots,\bar{T}$ in this limit set. We identify the parameters of $Y_d(t,X)$, $t=1,\ldots,\bar{T}$, and $M_d(X)$ only up to scale normalizations. We know the limit set given the functional forms for the $\Psi(t,Z)$ used in Theorem 1 or in Matzkin (1992, 1993, 1994).

As a consequence of *(ii)*, we can identify $\mu_c(t,X)$, $\mu_{c,M}(X)$ directly from the means of the limit outcome distributions. We can thus identify all pairwise average treatment effects $E(Y_c(t,X) \mid X=x) - E(Y_c(t',X) \mid X=x)$ for all t,t' and any other linear functionals derived from the distributions of the continuous variables defined at t and t'. Identification of the means and distributions of the latent variables giving rise to the discrete outcomes is more subtle. The argument required is the same as that used in the first step of the proof of Theorem 1. With one continuous regressor among the X, one can identify the marginal distributions of the $U_d(t)$ and the $U_{d,M}$ (up to scale if the Matzkin functions are only specified

up to scale). To identify the joint distributions of $U_d(t)$ and $U_{d,M}$ one must invoke a version of condition *(iii)* used in the proof of Theorem 1.

Thus for system t, suppose that there are $N_{d,t}$ discrete outcome components with associated means $\mu_{d,j}(t,X)$ and error terms $U_{d,j}(t)$, $j=1,\ldots,N_{d,t}$. As a consequence of condition *(iii)* of this Theorem, $Supp(\mu_d(t,X)) = Supp(\mu_{d,1}(t,X)) \times \cdots \times Supp\left(\mu_{d,N_{d,t}}(t,X)\right)$ and $Supp(\mu_d(t,X)) \supseteq Supp(U_d(t))$. We thus can trace out the joint distribution of $U_d(t)$ and identify it (up to scale if we specify the Matzkin class only up to scale). By a parallel argument for the measurements, we can identify the joint distribution of $U_{d,M}$. Let $N_{d,M}$ be the number of discrete measurements. From condition *(iii)*, we obtain $Supp\left(\mu_{d,M}(X)\right) = Supp(\mu_{d,M,1}(X)) \times \cdots \times Supp\left(\mu_{d,M,N_{d,M}}(X)\right)$ and $Supp\left(\mu_{d,M}(X)\right) \supseteq Supp(U_{d,M})$. Under these conditions, we can trace out the joint distribution of $U_{d,M}$ and identify it (up to scale for Matzkin class of functions specified up to scale) within the limit sets.

In the general case, we can vary each limit of the integral in (D.1) independently and trace out the full joint distribution of $(U_c(t), U_d(t), U_{c,M}, U_{d,M}, \eta(1), \ldots, \eta(t))$. For further discussion, see the analysis in CHH, Theorem 3. ■

Acknowledgements

This paper previously circulated under the title "Dynamic Treatment Effects." This research was supported by NSF SES-0099195, SES-0241858 and NIH R01-HD043411. Versions of this paper were presented at the Summer 1998 North American Meetings of the Econometric Society, the 1998 Canadian Econometric Study Group at the University of Western Ontario, the Midwest Econometrics Group in October 2000, the UCLA conference on Panel Data in April 2004, the Econometrics Study Group, UCL, London in June 2004, the Econometrics seminars at the University of Toulouse in November 2004, at Northwestern University in April 2005 and at the University of California at Berkeley in May 2005. We are grateful to the editor, Steve Durlauf, and an anonymous referee, as well as Xiaohong Chen, Jean-Pierre Florens, Han Hong, Weerachart Kilenthong, Thierry Magnac, John Rust, Mohan Singh, Jora Stixrud, Chris Taber, Petra Todd and especially Jaap Abbring, Flavio Cunha, Rosa Matzkin, Sergio Urzua and Edward Vytlacil for comments on various drafts of this paper. A website at http://jenni.uchicago.edu/dyn-trmt-eff contains supplementary material on the proofs reported in this paper.

References

Aakvik, A., J. J. Heckman, and E. J. Vytlacil (2005). Estimating treatment effects for discrete outcomes when responses to treatment vary: An application to Norwegian vocational rehabilitation programs. *Journal of Econometrics 125*(1-2), 15–51.

Abbring, J. (2000, July). The non-parametric identification of mixed semi Markov event history models. Vrije University, Amsterdam.

Abbring, J. and J. J. Heckman (2006). Dynamic policy evaluation. In L. Matyas and P. Sevestre (Eds.), *The Econometrics of Panel Data*. Kluwer Academic Publishers.

Abbring, J. H. and G. J. Van Den Berg (2003, September). The nonparametric identification of treatment effects in duration models. *Econometrica 71*(5), 1491–1517.

Aguirregabiria, V. (2004). Identification and estimation of dynamic input demand models: A discrete choice approach. Presented at Univesity of Chicago Department of Economics Seminar, December 2, 2004.

Anderson, T. and H. Rubin (1956). Statistical inference in factor analysis. In J. Neyman (Ed.), *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, 5*, pp. 111–150. Berkeley: University of California Press.

Belzil, C. and J. Hansen (2002, September). Unobserved ability and the return to schooling. *Econometrica 70*(5), 2075–2091.

Bonhomme, S. and J.-M. Robin (2004). Nonparametric identification and estimation of independent factor models. Unpublished working paper, Sorbonne, Paris.

Cameron, S. V. and J. J. Heckman (1998, April). Life cycle schooling and dynamic selection bias: Models and evidence for five cohorts of American males. *Journal of Political Economy 106*(2), 262–333.

Carneiro, P., K. Hansen, and J. J. Heckman (2001, Fall). Removing the veil of ignorance in assessing the distributional impacts of social policies. *Swedish Economic Policy Review 8*(2), 273–301.

Carneiro, P., K. Hansen, and J. J. Heckman (2003, May). Estimating distributions of treatment effects with an application to the returns to schooling and measurement of the effects of uncertainty on college choice. *International Economic Review 44*(2), 361–422. 2001 Lawrence R. Klein Lecture.

Cunha, F., J. J. Heckman, and S. Navarro (2005a). Counterfactual analysis of inequality and social mobility. In S. L. Morgan, D. B. Grusky, and G. S. Fields (Eds.), *Mobility and*

Inequality: Frontiers of Research from Sociology and Economics, Chapter 4. Palo Alto: Stanford University Press. forthcoming.

Cunha, F., J. J. Heckman, and S. Navarro (2005b, August). The evolution of uncertainty in the US economy. Presented at the 9th World Congress of the Econometric Society, London. Previously "Separating Heterogeneity from Uncertainty in an Aiyagari-Laitner Economy," presented at the Goldwater Conference on Labor Markets, Arizona State University, March 2004.

Cunha, F., J. J. Heckman, and S. Navarro (2005c). A framework for the analysis of inequality. *Journal of Macroeconomics*, forthcoming.

Cunha, F., J. J. Heckman, and S. Navarro (2005d). The identification and economic content of ordered choice models with stochastic cutoffs. Unpublished manuscript, University of Chicago, Department of Economics.

Cunha, F., J. J. Heckman, and S. Navarro (2005e, April). Separating uncertainty from heterogeneity in life cycle earnings, the 2004 Hicks Lecture. *Oxford Economic Papers 57*(2), 191–261.

Eckstein, Z. and K. I. Wolpin (1999, November). Why youths drop out of high school: The impact of preferences, opportunities, and abilities. *Econometrica 67*(6), 1295–1339.

Falmagne, J.-C. (1985). *Elements of Psychophysical Theory*. Oxford Psychology Series No. 6. New York: Oxford University Press.

Flinn, C. and J. J. Heckman (1982, January). New methods for analyzing structural models of labor force dynamics. *Journal of Econometrics 18*(1), 115–68.

Florens, J.-P., M. Mouchart, and J. Rolin (1990). *Elements of Bayesian Statistics*. New York: M. Dekker.

Gill, R. D. and J. M. Robins (2001, December). Causal inference for complex longitudinal data: The continuous case. *The Annals of Statistics 29*(6), 1785–1811.

Hansen, K. T., J. J. Heckman, and K. J. Mullen (2004, July-August). The effect of schooling and ability on achievement test scores. *Journal of Econometrics 121*(1-2), 39–98.

Heckman, J. J. (1974, July). Shadow prices, market wages, and labor supply. *Econometrica 42*(4), 679–694.

Heckman, J. J. (1981a). Heterogeneity and state dependence. In S. Rosen (Ed.), *Studies in Labor Markets*, National Bureau of Economic Research, pp. 91–139. University of Chicago Press.

Heckman, J. J. (1981b). The incidental parameters problem and the problem of initial conditions in estimating a discrete time-discrete data stochastic process and some Monte Carlo evidence. In C. Manski and D. McFadden (Eds.), *Structural Analysis of Discrete Data with Econometric Applications*, pp. 179–85. Cambridge, MA: MIT Press.

Heckman, J. J. (1981c). Statistical models for discrete panel data. In C. Manski and D. McFadden (Eds.), *Structural Analysis of Discrete Data with Econometric Applications*, pp. 114–178. Cambridge, MA: MIT Press.

Heckman, J. J. (1990, May). Varieties of selection bias. *American Economic Review 80*(2), 313–318.

Heckman, J. J. (2006). The scientific model of causality. *Sociological Methodology*, forthcoming.

Heckman, J. J. and G. J. Borjas (1980, August). Does unemployment cause future unemployment? Definitions, questions and answers from a continuous time model of heterogeneity and state dependence. *Economica 47*(187), 247–283. Special Issue on Unemployment.

Heckman, J. J. and B. E. Honoré (1989, June). The identifiability of the competing risks model. *Biometrika 76*(2), 325–330.

Heckman, J. J. and B. E. Honoré (1990, September). The empirical content of the Roy model. *Econometrica 58*(5), 1121–1149.

Heckman, J. J., R. J. LaLonde, and J. A. Smith (1999). The economics and econometrics of active labor market programs. In O. Ashenfelter and D. Card (Eds.), *Handbook of Labor Economics*, Volume 3A, Chapter 31, pp. 1865–2097. New York: North-Holland.

Heckman, J. J., L. J. Lochner, and C. Taber (1998, January). Explaining rising wage inequality: Explorations with a dynamic general equilibrium model of labor earnings with heterogeneous agents. *Review of Economic Dynamics 1*(1), 1–58.

Heckman, J. J., L. J. Lochner, and P. E. Todd (2006). Earnings equations and rates of return: The Mincer equation and beyond. In E. A. Hanushek and F. Welch (Eds.), *Handbook of the Economics of Education*. Amsterdam: North-Holland. forthcoming.

Heckman, J. J. and T. E. MaCurdy (1980, January). A life cycle model of female labour supply. *Review of Economic Studies 47*(1), 47–74.

Heckman, J. J. and S. Navarro (2004, February). Using matching, instrumental variables, and control functions to estimate economic choice models. *Review of Economics and Statistics 86*(1), 30–57.

Heckman, J. J. and S. Navarro (2006). Empirical estimates of option values of education and information sets in a dynamic sequential choice model. Unpublished manuscript, University of Chicago, Department of Economics.

Heckman, J. J. and B. S. Singer (1984, March). A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica 52*(2), 271–320.

Heckman, J. J. and J. A. Smith (1998). Evaluating the welfare state. In S. Strom (Ed.), *Econometrics and Economic Theory in the Twentieth Century: The Ragnar Frisch Centennial Symposium*, pp. 241–318. New York: Cambridge University Press.

Heckman, J. J., J. Stixrud, and S. Urzua (2004). Noncognitive skills. Unpublished manuscript, University of Chicago, Department of Economics.

Heckman, J. J., J. L. Tobias, and E. J. Vytlacil (2001, October). Four parameters of interest in the evaluation of social programs. *Southern Economic Journal 68*(2), 210–223.

Heckman, J. J., J. L. Tobias, and E. J. Vytlacil (2003, August). Simple estimators for treatment parameters in a latent variable framework. *Review of Economics and Statistics 85*(3), 748–754.

Heckman, J. J., S. Urzua, and E. J. Vytlacil (2005). Understanding instrumental variables in models with essential heterogeneity. *Review of Economics and Statistics* Lecture, 2002.

Heckman, J. J., S. Urzua, and G. Yates (2005). The identification and estimation of option values in a model with recurrent states. Unpublished manuscript, University of Chicago, Department of Economics.

Heckman, J. J. and E. J. Vytlacil (1999, April). Local instrumental variables and latent variable models for identifying and bounding treatment effects. *Proceedings of the National Academy of Sciences 96*, 4730–4734.

Heckman, J. J. and E. J. Vytlacil (2001). Causal parameters, treatment effects and randomization. Unpublished manuscript, University of Chicago, Department of Economics.

Heckman, J. J. and E. J. Vytlacil (2005, May). Structural equations, treatment effects and econometric policy evaluation. *Econometrica 73*(3), 669–738.

Heckman, J. J. and E. J. Vytlacil (2006a). Econometric evaluation of social programs, Part I: Causal models, structural models and econometric policy evaluation. In J. Heckman

and E. Leamer (Eds.), *Handbook of Econometrics, Volume 6*. Amsterdam: Elsevier. forthcoming.

Heckman, J. J. and E. J. Vytlacil (2006b). Econometric evaluation of social programs, Part II: Using economic choice theory and the marginal treatment effect to organize alternative econometric estimators. In J. Heckman and E. Leamer (Eds.), *Handbook of Econometrics, Volume 6*. Amsterdam: Elsevier. forthcoming.

Holland, P. W. (1986, December). Statistics and causal inference. *Journal of the American Statistical Association 81*(396), 945–960.

Horowitz, J. L. (1998). *Semiparametric Methods in Econometrics*. New York: Springer.

Hotz, V. J. and R. A. Miller (1988, January). An empirical analysis of life cycle fertility and female labor supply. *Econometrica 56*(1), 91–118.

Hotz, V. J. and R. A. Miller (1993, July). Conditional choice probabilities and the estimation of dynamic models. *Review of Economic Studies 60*(3), 497–529.

Keane, M. P. and K. I. Wolpin (1994, November). The solution and estimation of discrete choice dynamic programming models by simulation and interpolation: Monte carlo evidence. *The Review of Economics and Statistics 76*(4), 648–672.

Keane, M. P. and K. I. Wolpin (1997, June). The career decisions of young men. *Journal of Political Economy 105*(3), 473–522.

Lechner, M. and R. Miquel (2002). Identification of effects of dynamic treatments by sequential conditional independence assumptions. Discussion paper, University of St. Gallen, Department of Economics.

Lewbel, A. (2000, July). Semiparametric qualitative response model estimation with unknown heteroscedasticity or instrumental variables. *Journal of Econometrics 97*(1), 145–177.

Lok, J. J. (2001). *Statistical Modelling of Causal Effects in Time*. Ph. D. thesis, Free University, Amsterdam. Division of Mathematics and Computer Science, Faculty of Sciences,.

MaCurdy, T. E. (1981, December). An empirical model of labor supply in a life-cycle setting. *Journal of Political Economy 89*(6), 1059–1085.

Magnac, T. and E. Maurin (2005). Identification and information in monotone binary models. *Journal of Econometrics*, forthcoming.

Magnac, T. and D. Thesmar (2002, March). Identifying dynamic discrete decision processes. *Econometrica 70*(2), 801–816.

Manski, C. F. (1988, September). Identification of binary response models. *Journal of the American Statistical Association 83*(403), 729–738.

Manski, C. F. (1993, July). Dynamic choice in social settings: Learning from the experiences of others. *Journal of Econometrics 58*(1-2), 121–136.

Manski, C. F. (2003). *Partial Identification of Probability Distributions*. New York: Springer-Verlag.

Manski, C. F. (2004, September). Measuring expectations. *Econometrica 72*(5), 1329–1376.

Matzkin, R. L. (1992, March). Nonparametric and distribution-free estimation of the binary threshold crossing and the binary choice models. *Econometrica 60*(2), 239–270.

Matzkin, R. L. (1993, July). Nonparametric identification and estimation of polychotomous choice models. *Journal of Econometrics 58*(1-2), 137–168.

Matzkin, R. L. (1994). Restrictions of economic theory in nonparametric methods. In R. Engle and D. McFadden (Eds.), *Handbook of Econometrics*, Volume 4, pp. 2523–58. New York: North-Holland.

Matzkin, R. L. (2003, September). Nonparametric estimation of nonadditive random functions. *Econometrica 71*(5), 1339–1375.

Miller, R. A. (1984, December). Job matching and occupational choice. *Journal of Political Economy 92*(6), 1086–1120.

Mincer, J. (1974). *Schooling, Experience and Earnings*. New York: National Bureau of Economic Research.

Navarro, S. (2004a). Semiparametric identification of factor models for counterfactual analysis. Unpublished manuscript, University of Chicago, Department of Economics.

Navarro, S. (2004b). Understanding schooling: Using observed choices to infer agent's information in a dynamic model of schooling choice when consumption allocation is subject to borrowing constraints. Unpublished manuscript, University of Chicago, Department of Economics.

Pakes, A. (1986, July). Patents as options: Some estimates of the value of holding European patent stocks. *Econometrica 54*(4), 755–784.

Pakes, A. and M. Simpson (1989). Patent renewal data. *Brookings Papers on Economic Activity* (Special Issue), 331–401.

Powell, J. L., J. H. Stock, and T. M. Stoker (1989, November). Semiparametric estimation of index coefficients. *Econometrica 57*(6), 1403–1430.

Ridder, G. (1990, April). The non-parametric identification of generalized accelerated failure-time models. *Review of Economic Studies 57*(2), 167–181.

Robins, J. M. (1989). The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies. In L. Sechrest, H. Freeman, and A. Mulley (Eds.), *Health Services Research Methodology: A Focus on*

AIDS, pp. 113–159. Rockville, MD: U.S. Department of Health and Human Services, National Center for Health Services Research and Health Care Technology Assessment.

Robins, J. M. (1997). Causal inference from complex longitudinal data. In M. Berkane (Ed.), *Latent Variable Modeling and Applications to Causality. Lecture Notes in Statistics*, pp. 69–117. New York: Springer-Verlag.

Rosenbaum, P. R. and D. B. Rubin (1983, April). The central role of the propensity score in observational studies for causal effects. *Biometrika 70*(1), 41–55.

Rust, J. (1987, September). Optimal replacement of GMC bus engines: An empirical model of Harold Zurcher. *Econometrica 55*(5), 999–1033.

Rust, J. (1994). Structural estimation of Markov decision processes. In R. Engle and D. McFadden (Eds.), *Handbook of Econometrics, Volume*, pp. 3081–3143. New York: North-Holland.

Taber, C. R. (2000, June). Semiparametric identification and heterogeneity in discrete choice dynamic programming models. *Journal of Econometrics 96*(2), 201–229.

Thurstone, L. L. (1959). *The Measurement of Values*. Chicago: University of Chicago Press.

Urzua, S. (2005). Schooling choice and the anticipation of labor market conditions: A dynamic choice model with heterogeneous agents and learning. Unpublished manuscript, University of Chicago, Department of Economics.

Van der Laan, M. J. and J. M. Robins (2003). *Unified Methods for Censored Longitudinal Data and Causality*. New York: Springer-Verlag.

Wolpin, K. I. (1984, October). An estimable dynamic stochastic model of fertility and child mortality. *Journal of Political Economy 92*(5), 852–874.

Wolpin, K. I. (1987, July). Estimating a structural search model: The transition from school to work. *Econometrica 55*(4), 801–817.

Figure 1. Evolution of grades and age

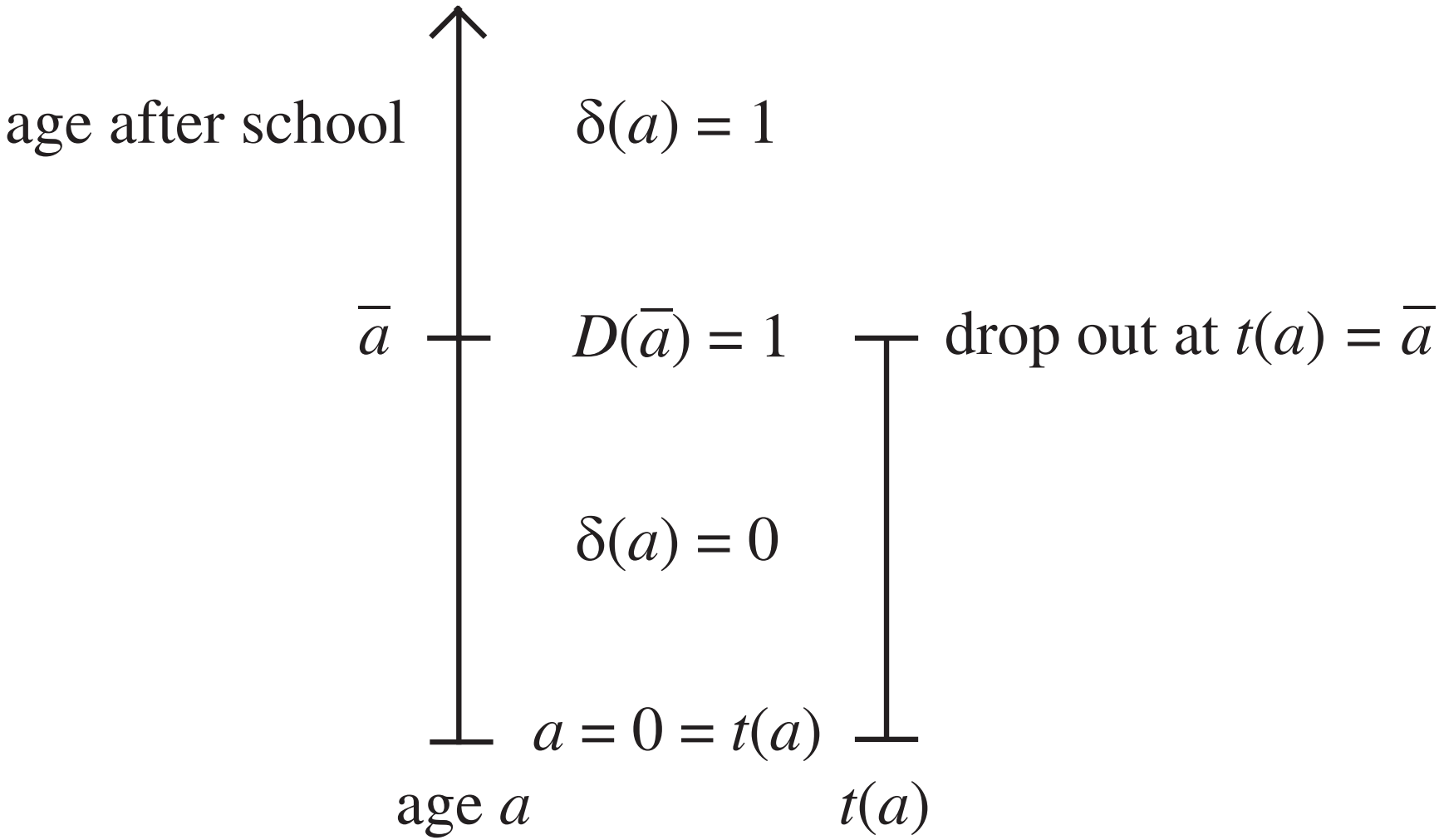
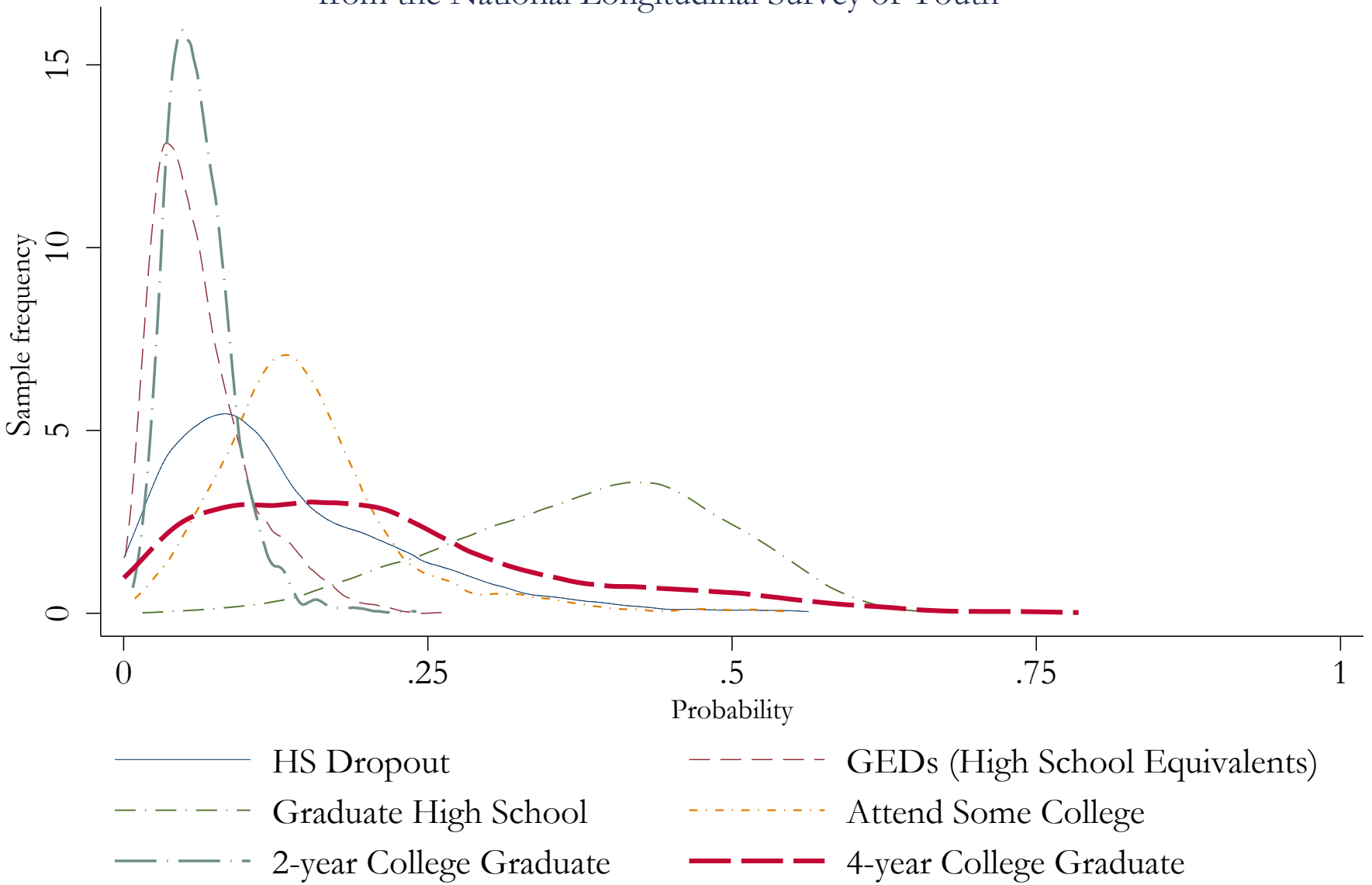


Figure 2. Sample distribution of schooling attainment probabilities for males from the National Longitudinal Survey of Youth



Source: Heckman, Stixrud and Urzua (2005)