WORKING PAPER · NO. 2019-11

Nonparametric Inference on State Dependence in Unemployment

Alexander Torgovitsky

JANUARY 2019

Becker Friedman Institute
FOR ECONOMICS AT UCHICAGO

1126 E. 59th St, Chicago, IL 60637
Main: 773.702.5599
bfi.uchicago.edu

Nonparametric Inference on State Dependence in Unemployment*

Alexander Torgovitsky†

January 2, 2019

Abstract

This paper is about measuring state dependence in dynamic discrete outcomes. I develop a nonparametric dynamic potential outcomes (DPO) model and propose an array of parameters and identifying assumptions that can be considered in this model. I show how to construct sharp identified sets under combinations of identifying assumptions by using a flexible linear programming procedure. I apply the analysis to study state dependence in unemployment for working age high school educated men using an extract from the 2008 Survey of Income and Program Participation (SIPP). Using only nonparametric assumptions, I estimate that state dependence accounts for at least 30–40% of the four-month persistence in unemployment among high school educated men.

JEL classification: C14; C20; C51; J2; J6
Keywords: State dependence, unemployment, nonparametric, partial identification, linear programming, dynamic discrete choice, moment inequalities

*The first draft of this paper was circulated under the title "Partial Identification of State Dependence" and dated February 12, 2015. The second draft was circulated under the title "Nonparametric Inference on State Dependence with Applications to Employment Dynamics" and dated January 29, 2016.

†Department of Economics, University of Chicago. This paper was presented at the University of Wisconsin at Madison, Northwestern University, the University of Chicago, Columbia University, the 2015 World Congress of the Econometric Society, the 2016 North American Winter Meeting of the Econometric Society, and the 2016 East Asian Meeting of the Econometric Society. My thanks to the audiences at those presentations. I thank Ivan Canay, Andres Santos and Azeem Shaikh for several helpful discussions. Useful comments and feedback were provided by Stéphane Bonhomme, Xiaohong Chen, Iván Fernández-Val, Joachim Freyberger, Jim Heckman, Joel Horowitz, Hide Ichimura, Yuichi Kitamura, Pat Kline, Thibaut Lamadon, Chuck Manski, Whitney Newey, Matt Notowidigdo, Jack Porter, Nancy Stokey, and Tiemen Woutersen. Four anonymous referees and the co-editor provided constructive comments that led to substantial improvements in the paper. This research was supported in part by National Science Foundation grant SES-1530538. I am grateful to the Department of Economics at the University of Chicago for supporting me hospitably through a visiting position while I completed part of this research in 2015–2016.

1 Introduction

Suppose that a researcher observes a balanced panel consisting of a binary outcome $Y_{it} \in \{0,1\}$ at time periods $t = 0, 1, \ldots, T$ for a cross-section of agents i. The researcher's goal is to determine to what extent the outcome in the previous period, $Y_{i(t-1)}$, has a causal effect on the current period outcome, Y_{it}. For example, Heckman (1981a) studied whether past employment has a causal effect on future employment for married women. A negative causal effect could arise from search costs, human capital depreciation during non-employment, or quality signaling in hiring processes ("stigma" or "scarring" effects), among other explanations. Such an effect is commonly described as state dependence, or "true" state dependence for emphasis.

Positive serial correlation in employment outcomes $Y_i \equiv (Y_{i0}, Y_{i1}, \ldots, Y_{iT})$ does not necessarily indicate state dependence. An alternative explanation is that individuals have persistent latent heterogeneity in their propensities for employment and, as a result, some individuals are always more likely to be employed than others (Heckman and Willis, 1977; Heckman, 1978, 1981a). This would lead to positive serial correlation in observed employment outcomes even if there is no state dependence in employment.

The difference between these two explanations has important implications for the design and long-run efficacy of active labor market programs (Heckman, 1978, 1981a,b). It is therefore important to have convincing econometric methods to quantify the degree to which persistence in employment is due to state dependence. In order to be convincing, these econometric methods must first address the difficult identification problem of distinguishing state dependence from persistent unobserved heterogeneity.

The main contribution of this paper is the development of a new nonparametric framework for measuring state dependence. The framework is a dynamic potential outcomes (DPO) model, the premise of which is simple to state. Given a binary outcome $Y_{it} \in \{0,1\}$ for agent i at time t, let $U_{it}(0)$ and $U_{it}(1)$ be two latent binary variables that represent the potential outcomes that would have been realized had the prior period outcome, $Y_{i(t-1)}$, counterfactually been 0 or 1, respectively. The observed outcome is therefore related to the potential outcomes as

$$Y_{it} = Y_{i(t-1)} U_{it}(1) + (1 - Y_{i(t-1)}) U_{it}(0).$$

The model primitive is the joint distribution of $U_{it}(0), U_{it}(1)$ across all time periods.[1]

[1] After reading an early draft of this paper, Chuck Manski shared with me his slides for an invited talk in 2006 in which he proposed using the same type of model to study state dependence (Manski, 2006). This paper was developed independently and without knowledge of that talk. The analysis of the DPO model in this paper is significantly different than that in Manski's talk.

This model can be used to construct a number of interesting measures of state dependence, including common measures such as the average treatment effect. I discuss several parameters that provide different measures of state dependence, and I show that they are usually not point identified. For three of them, I derive sharp worst-case bounds that use only the empirical evidence. The bounds are quite wide. In particular, the bounds imply that empirical evidence alone is never informative enough to reject the hypothesis of no state dependence. At the same time, the empirical evidence alone is also never informative enough to reject the hypothesis that *all* of the observed persistence in the data is due to state dependence.

I propose several additional nonparametric assumptions that can be maintained for more informative inference. The assumptions concern the temporal dependence and stationarity of the potential outcomes, as well as their relationships with other observed covariates. These assumptions are fully nonparametric and have intuitive interpretations. For additional interpretation, I consider the DPO model that is implied by a dynamic model of a forward-looking agent making discrete choices. I develop nonparametric conditions on the choice model that are sufficient to imply each of the conditions on the DPO model. I also consider specializations of these conditions to a widely used dynamic binary response model.

Since the DPO model is recursive, analytically deriving sharp bounds under additional assumptions is quite difficult. Instead, I develop a general procedure for computing sharp bounds that is valid for broad classes of parameters and combinations of assumptions. In many cases, the procedure amounts to solving two linear programming problems and is therefore straightforward to implement. An attractive feature of this approach is its flexibility: The researcher is afforded greater freedom to choose parameters and combine assumptions, without needing to derive new analytic results for each new specification.

The econometric methodology proposed in this paper can be applied to any of the large variety of empirical settings in which identifying state dependence is important. These include the dynamics of welfare recipiency (Chay, Hoynes, and Hyslop, 2004; Card and Hyslop, 2005), product choices among consumers (Keane, 1997; Dubé, Hitsch, and Rossi, 2010; Handel, 2013), self-reported health status (Contoyannis, Jones, and Rice, 2004), firm investment (Drakos and Konstantinou, 2013) and exporting (Bernard and Jensen, 2004) decisions, household investment behavior (Alessie, Hochguertel, and Soest, 2004), illicit drug usage (Deza, 2015), and eating disorders (Ham, Iorio, and Sovinsky, 2013). Irace (2018) used the methodology developed in this paper to study the dynamics of hospital choice.

I apply the methodology to study the employment dynamics of working age, high

school educated men, using an extract from the 2008 Survey of Income and Program Participation (SIPP) that covers January 2011 to April 2013. I find little evidence of state dependence among employed workers. However, by maintaining a nonparametric stationarity assumption, I find evidence of substantial state dependence among unemployed workers. The main estimates indicate that at least 23% of unemployed workers would be employed if and only if they had been employed in the previous period. Overall, state dependence accounts for at least 41% of the observed four-month persistence in unemployment. The results imply that short-term state dependence is an important phenomenon in the U.S. labor market.

The organization of this paper is as follows. In the next section, I develop the DPO model and connect it to a dynamic discrete choice model. In Section 3, I discuss parameters of interest in the DPO model, derive worst-case bounds, and develop a general procedure for computing sharp identified sets. In Section 4, I propose an array of identifying assumptions that can be imposed in the DPO model. I analyze the economic content of these assumptions through the lens of the dynamic discrete choice model. In Section 5, I apply the DPO model to study state dependence in unemployment. Section 6 contains a brief conclusion.

2 The Dynamic Potential Outcomes Model

2.1 Model

The canonical *static* potential outcomes model is based on two unobserved outcomes, $U_i(0)$ and $U_i(1)$, that would have been obtained had a binary treatment, $D_i \in \{0,1\}$, been exogenously manipulated to be 0 or 1. The observed outcome, Y_i, is related to the potential outcomes and the observed treatment state through $Y_i = D_i U_i(1) + (1 - D_i)U_i(0)$. The researcher is interested in inferring features of the unobservable distribution of $(U_i(0), U_i(1))$ from the observable distribution of (Y_i, D_i).

State dependence is the causal effect of a past outcome on a current outcome. At time t, the outcome is the current outcome Y_{it}, and the "treatment" is the immediately preceding outcome, $Y_{i(t-1)}$.[2] I assume throughout the main text that $Y_{it} \in \{0,1\}$ is binary for each t and discuss the extension to multi-valued outcomes in Appendix A. Thus, in analogy to the static potential outcomes model, suppose that for each time period $t = 1, \ldots, T$ there exists unobservable random variables $U_{it}(0)$ and $U_{it}(1)$ taking values in $\{0,1\}$. These binary unobservables represent the outcome that would have been realized at time t had the past period outcome $Y_{i(t-1)}$ been exogenously

[2]Note in particular the distinction with the dynamic treatment effects literature (e.g. Abbring and Heckman, 2007; Angrist and Kuersteiner, 2011), in which the treatment and outcome variables are distinct.

manipulated to be 0 or 1, respectively.

The observed outcomes $Y_i \equiv (Y_{i0}, Y_{i1}, \ldots, Y_{iT})$ together form a random vector with values in $\mathcal{Y} \equiv \{0,1\}^{T+1}$, the $(T+1)$–fold Cartesian product of $\{0,1\}$. The observed outcomes are related to potential outcomes $U_i(0) \equiv (U_{i1}(0), \ldots, U_{iT}(0))$ and $U_i(1) \equiv (U_{i1}(1), \ldots, U_{iT}(1))$ through the recursive relationship

$$Y_{it} = Y_{i(t-1)}U_{it}(1) + (1 - Y_{i(t-1)})U_{it}(0) = U_{it}(Y_{i(t-1)}) \quad \text{for all } t \geq 1. \tag{1}$$

In this formulation, the outcome in the initial period, Y_{i0}, is observed but not modeled. This avoids the initial conditions problem discussed by Heckman (1981c) by simply reducing the number of observed variables that are explicitly modeled, similar in spirit to the approach of Honoré and Tamer (2006) or Chen, Tamer, and Torgovitsky (2011).[3]

This specification presumes that the researcher is only interested in the causal effect of the outcome in the immediately preceding period on the outcome in the current period. In some settings, it may be interesting to analyze the causal effects of longer sequences of prior outcomes on the current period outcome. This can be accommodated by redefining the potential outcomes to include a separate potential outcome for every sequence up to a certain length. For clarity, I focus on the one-period causal effect in the main text and discuss this extension to longer sequences in Appendix B. However, note that focusing on single period sequences in (1) does not place any restrictions on the temporal dependence of the potential outcomes. In particular, even though only first-order causal effects are being modeled, (1) *does not* imply that the potential outcomes follow a first-order Markov chain.

In addition to Y_i, the researcher also observes a vector $X_i = (X_{i0}, X_{i1}, \ldots, X_{iT})$ of covariates with support $\mathcal{X}$. The components of X_{it} may be time-varying or time-invariant. I assume for simplicity that $\mathcal{X}$ is a finite set, so that X_i is discretely distributed.[4] Some of the components of X_{it} may be thought of as conditioning variables that describe observed heterogeneity, while others might be viewed as instruments that satisfy certain exclusion or monotonicity conditions. These types of assumptions are discussed in Section 4.5.

The DPO model captures state dependence through the possibility that $U_{it}(0) \neq U_{it}(1)$. That is, the outcome $Y_{it} = U_{it}(Y_{i(t-1)})$ that actually occurred for agent i in

[3]In particular, note that the DPO model *does not* impose independence between Y_{i0} and any of the subsequent potential outcomes. It is straightforward to add such a condition as an additional identifying assumption, but this is often difficult to justify (Heckman, 1981c), so I do not consider it in this paper.

[4]Continuous covariates do not present any conceptual difficulty for the identification analysis, see the discussion in e.g. Torgovitsky (forthcoming). However, as is usually the case in nonparametric analyses, they do complicate estimation and statistical inference, so for simplicity I focus on the discrete case.

period t may have been different had $Y_{i(t-1)}$ been different. The DPO model allows for "occurrence," "duration," and "lagged duration" dependence, as defined by Heckman and Borjas (1980). It also allows for general forms of both observed and unobserved heterogeneity. Observed heterogeneity is captured through differences in the distributions of $(U_i(0), U_i(1))|X_i = x$ for different values of x. Unobserved heterogeneity is captured through variation in $(U_i(0), U_i(1))$, conditional on $X_i = x$. For example, the model allows for the possibility that conditional on $X_i = x$, $U_{it}(1) - U_{it}(0)$ is a random variable taking values in $\{-1, 0, 1\}$ for agents that differ along unobservable characteristics such as preferences or private information. The basic DPO model does not separate this unobserved heterogeneity into persistent and transitory components, and so does not impose any restrictions on the serial dependence of the potential outcomes. In Section 4, I discuss several assumptions that can be imposed to create a permanent-transitory distinction.

2.2 DPO Models Implied by Dynamic Choice Models

In this section, I connect the DPO model to a discrete time dynamic choice (DC) model of a rational, forward-looking economic agent. This serves two purposes. First, the DC model will be used to motivate and interpret the additional identifying assumptions for the DPO model that are proposed in Section 4. Second, since the DC model nests standard "structural" and "reduced form" models as special cases, it provides a vehicle for comparing these models to the DPO model.

The DC model is as follows. Time runs from some initial period $\underline{T}$ that occurs at or before $t = 0$ to some terminal period $\bar{T}$ that occurs at or after T, where $\bar{T}$ may be either finite or infinite. In each period t, agent i chooses $C_{it} \equiv (Y_{it}, D_{it})$. One of these choice variables is the binary outcome, Y_{it}, that the researcher observes in periods $t = 0, 1, \ldots, T$. The other choice variables, D_{it}, could take any number of values, and could be either observed or unobserved by the researcher.

Agent i receives flow utility in period t of $\mu(Y_{it}, S_{it})$, where $S_{it} \equiv (C_{i(t-1)}, Z_{it})$ are state variables that may affect this utility. I assume throughout that the flow utility is bounded. The state variables include the previous period choices, $C_{i(t-1)}$, and an additional vector of exogenous state variables, Z_{it}, which could contain both observable and unobservable components.

Each agent maximizes their expected present-discounted utility using discount factor $\delta \in (0, 1)$.[5] Under mild regularity conditions (see, e.g. Stokey, Lucas, and Prescott,

[5]In all of the following, δ could be replaced by δ_i and allowed to vary over the population as long as $\delta_i \in (0, 1)$ with probability 1. In this case, one could treat δ_i as a component of Z_{it}.

1989 or Rust, 1994), agent i's problem can be written recursively in terms of the Bellman equation

$$\nu(S_{it}) = \max_{c' \in \mathcal{C}} \left\{ \mu(c', S_{it}) + \delta \int \nu(c', s') \, d\Lambda(s'|S_{it}) \right\} \equiv \max_{c' \in \mathcal{C}} \mathring{\nu}(c', S_{it}), \tag{2}$$

where ν is the value function, $\mathcal{C} \equiv \{0,1\} \times \mathcal{D}$ is the feasible set of choices, $\Lambda(\cdot|s)$ is a distribution function for $S_{i(t+1)}$, conditional on $S_{it} = s$, and $\mathring{\nu}$ is shorthand notation that combines the flow utility and continuation value. The distribution function, Λ, captures the agent's beliefs about the evolution of the state variables, including deterministic laws of motion. Neither μ nor Λ have an i or a t subscript because all observable and unobservable differences across agents and time are viewed notationally as part of the state variables, S_{it}. So, for example, one component of S_{it} could be the time period itself, t, which would allow for these functions to vary over time.

To compare the predictions of this model to those of the DPO model, I will isolate the Y_{it} choice. Profile (2) as

$$\nu(S_{it}) = \max_{y' \in \{0,1\}} \max_{d' \in \mathcal{D}} \mathring{\nu}(y', d', S_{it}). \tag{3}$$

Suppose that there is a unique solution to the inner problem in (3) given the (possibly suboptimal) choice of y', and denote it as

$$d^\star(S_{it} \| y') \equiv \arg\max_{d' \in \mathcal{D}} \mathring{\nu}(y', d', S_{it}). \tag{4}$$

Then (3) can be written as

$$\nu(S_{it}) = \max_{y' \in \{0,1\}} \mathring{\nu}\left(y', d^\star(S_{it} \| y'), S_{it}\right) \equiv \max_{y' \in \{0,1\}} \mathring{\nu}\left(S_{it} \| y'\right). \tag{5}$$

The observed binary choice, Y_{it}, is assumed to be the optimizer of (5):

$$Y_{it} = \arg\max_{y' \in \{0,1\}} \mathring{\nu}(S_{it} \| y') = \mathbb{1}\left[\Delta\mathring{\nu}(S_{it}) \geq 0\right], \tag{6}$$

where $\Delta\mathring{\nu}(S_{it}) \equiv \mathring{\nu}(S_{it}\|1) - \mathring{\nu}(S_{it}\|0)$, and ties are broken in favor of $Y_{it} = 1$.

Corresponding to the agent's actual choice, Y_{it}, are two counterfactual choices that they would have made in time t, had they actually chosen $y \in \{0,1\}$ in period $t-1$. The counterfactual entertained here is that the agent also re-optimizes their choice of d at time $t-1$, given their choice of $Y_{i(t-1)} = y$. Thus, instead of $D_{i(t-1)}$ they choose $d^\star(S_{i(t-1)} \| y)$. The other state variables, Z_{it}, are presumed to remain the same in both

counterfactual states.[6] Let $S_{it}(y) \equiv (y, d^{\star}(S_{i(t-1)} \| y), Z_{it})$ denote the state variables that would have been realized at time t had the agent chosen y in period $t-1$. Then the agent's choice in period t if they had chosen y in period $(t-1)$ can be written as

$$U_{it}(y) = \mathbb{1}\left[\Delta \mathring{\nu}(S_{it}(y)) \geq 0\right]. \tag{7}$$

Equation (7) shows that a DC model generates a DPO model.[7]

2.3 DPO Models Implied by Dynamic Binary Response Models

Dynamic binary response (DBR) models are commonly used for measuring state dependence.[8] A textbook version of the model (e.g. Wooldridge, 2010, Section 15.8.4) has the threshold-crossing equation

$$Y_{it} = \mathbb{1}\left[\beta_0 Y_{i(t-1)} + X_{it}'\beta_1 + A_i + V_{it} \geq 0\right] \quad \text{for } t \geq 1, \tag{8}$$

where (β_0, β_1) are unknown parameters, and the exogenous state variables $Z_{it} = (X_{it}, A_i, V_{it})$ consist of an observable component, X_{it}, an unobservable time-invariant component, A_i, and an unobservable time-varying component, V_{it}.[9] This model can be viewed as a special case of (6) in which $\Delta \mathring{\nu}(S_{it}) = \beta_0 Y_{i(t-1)} + X_{it}'\beta_1 + A_i + V_{it}$. The associated potential outcomes are special cases of (7):

$$U_{it}(y) = \mathbb{1}[\beta_0 y + X_{it}'\beta_1 + A_i + V_{it} \geq 0] \quad \text{for } y \in \{0,1\} \text{ and } t \geq 1. \tag{9}$$

Typical implementations of (8) maintain the assumption that X_i is independent of $(A_i, V_{i1}, \ldots, V_{iT})$, that A_i is normally distributed, and that V_{it} are normally (or logistically) distributed. However, there are also several known results concerning semi- and non-parametric modifications of this and similar models. These are surveyed in Appendix C. One criticism of (8)–(9) is that the relationship between the primitives

[6] This by itself is not restrictive since any components that change can be treated as part of D_{it}.

[7] Note that $U_{it}(y)$ as defined in (7) satisfies the law of motion (1), since $U_{it}(Y_{i(t-1)}) = \mathbb{1}\left[\Delta \mathring{\nu}(S_{it}(Y_{it})) \geq 0\right] = \mathbb{1}\left[\Delta \mathring{\nu}(S_{it}) \geq 0\right] = Y_{it}$.

[8] Most of the empirical papers listed in the introduction use some variety of DBR model. An early example is Heckman (1981a). Linear probability models are also occasionally used to analyze state dependence in binary outcomes (e.g. pp. 1265–1266 of Hyslop (1999)), however they have highly undesirable properties when viewed as causal models (Manski and Pepper 2009, pg. S210).

[9] Additionally, one must account for the determination of Y_{i0} to account for the initial conditions problem observed by Heckman (1981c). One popular way to do this, proposed by Wooldridge (2005), is to include the initial period outcome Y_{i0} as one of the observed state variables, and interpret inference as conditional on Y_{i0}.

of the choice problem (2) and the specification of $\Delta\mathring{\nu}$ in (9) can be obscure.[10]

2.4 Discussion

The attraction of the "structural" DC model (6) is that is it derived directly from a model of choice behavior. However, starting with Rust (1994), it has been recognized that additional parametric assumptions must be maintained in order to point identify primitive features of this model; see also Magnac and Thesmar (2002). These assumptions include finitely parameterized functional forms for the flow utility, μ, as well as parametric distributions for unobserved components of the exogenous state variables. It is also commonly assumed that the observed exogenous state variables are independent of the unobserved components, that δ is fixed at a value known to the researcher, that agents have knowledge of the distribution of the unobserved state variables and, frequently, that agents have perfect foresight or rational expectations over all observed exogenous state variables. Many of these assumptions are often questionable in applications.

Of particular concern are parametric assumptions about the distributions of unobservables. As Flinn and Heckman (1982, pg. 132) observed, "It [economic theory] is silent on the topic of the correct specification of functional forms for the distributions of unobservables." More bluntly, there is typically little economic rationale for assuming that an unobservable state variable follows a normal distribution as opposed to a logistic, Gumbel, or mixture of normals, among many other possibilities. Even advocates of structural modeling recognize parametric functional form restrictions as undesirable and "extra–theoretic" (e.g. Keane, Todd, and Wolpin, 2011, pg. 452).[11]

The DPO model is fully nonparametric, so does not suffer from this drawback. However, like the DBR model (8), the DPO model has a more tenuous connection with an underlying choice model. The approach taken in this paper is to view the DPO model as being generated by an underlying DC model through (7). In Section 4, I show that nonparametric assumptions on the DC model imply nonparametric assumptions on its generated DPO model. In Section 5, I use a specific job search model to motivate the assumptions used in the application to unemployment dynamics. While the DPO model can be used by itself without taking this view, the link back to the familiar DC

[10] Although, see Hyslop (1999), who develops a choice model that approximately implies (8)–(9).

[11] Norets and Tang (2014) have made progress on rigorously characterizing the effect of such assumptions on identification for DC models. However, a completely distribution-free characterization remains a difficult problem even in static discrete choice models (Torgovitsky, forthcoming). See also Heckman and Navarro (2007), who consider point identification in nonparametric dynamic choice models using exogenous variables with large support. More semi- and nonparametric results are available for DBR models like (9); see Appendix C.

model can be helpful for interpretation.

3 Identification

3.1 Definitions

In this section, I develop a general procedure for constructing identified sets in the DPO model. The procedure is abstract with respect to the assumptions and parameter of interest; concrete examples are discussed ahead. I assume throughout the analysis that the panel is balanced.[12] Periods are indexed by $t = 0, 1, \ldots, T$ for T small and fixed, and probabilities are taken over agents i drawn from the population.

The DPO model is (1). The primitive of the DPO model is a probability mass function P with support contained in $\mathcal{U} \times \mathcal{X}$, where $\mathcal{U} \equiv \{0,1\}^{2T+1}$ is the collection of all possible realizations of $U_i \equiv (Y_{i0}, U_i(0), U_i(1))$. A function P with domain $\mathcal{U} \times \mathcal{X}$ is a probability mass function on $\mathcal{U} \times \mathcal{X}$ if and only if it takes values in $[0,1]$ and

$$\sum_{u \in \mathcal{U}, x \in \mathcal{X}} P(u,x) = 1. \tag{10}$$

Let $\mathcal{P}$ denote the set of all functions $P : \mathcal{U} \times \mathcal{X} \to [0,1]$ that satisfy (10).

The *parameter space*, $\mathcal{P}^\dagger$, is the subset of $\mathcal{P}$ that satisfies the researcher's prior assumptions. Notationally, it is convenient to describe $\mathcal{P}^\dagger$ as $\mathcal{P}^\dagger = \{P \in \mathcal{P} : \rho(P) \geq 0\}$, where $\rho : \mathcal{P} \to \mathbb{R}^{d_\rho}$ is a function representing restrictions on P, and the inequality is interpreted component-wise. Equality restrictions can be incorporated into $\mathcal{P}^\dagger$ by including pairs of inequalities in the function ρ. The restrictions may depend on features of the observable distribution of (Y_i, X_i), but this is suppressed in the notation.

The *identified set*, $\mathcal{P}^\star$, is defined as the subset of $\mathcal{P}^\dagger$ that could have generated the observed data through relationship (1). Let $\mathbb{P}[Y_i = \cdot, X_i = \cdot]$ denote the observable probability mass function of (Y_i, X_i), where $Y_i \equiv (Y_{i0}, Y_{i1}, \ldots, Y_{iT})$. Then $P \in \mathcal{P}^\star$ requires that for every $y \equiv (y_0, y_1, \ldots, y_T) \in \mathcal{Y}$ and $x \in \mathcal{X}$,

$$\begin{aligned}\mathbb{P}[Y_i = y, X_i = x] &= \mathbb{P}_P[Y_i = y, X_i = x] \\ &= \mathbb{P}_P\left[Y_{i0} = y_0, U_{it}(y_{t-1}) = y_t \text{ all } t \geq 1, X_i = x\right],\end{aligned}$$

where $\mathbb{P}_P[\cdot]$ denotes the probability of an event when (U_i, X_i) is distributed according to P and Y_i is determined recursively through (1). This expression can be rewritten

[12] The analysis can be extended to unbalanced panels with either random or non-random attrition by conditioning on the time period at which attrition occurs.

Figure 1: Observational Equivalence, $T = 2$

Potential Outcomes					Observed Outcomes		
Y_{i0}	$U_{i1}(0)$	$U_{i1}(1)$	$U_{i2}(0)$	$U_{i2}(1)$	Y_{i0}	Y_{i1}	Y_{i2}
$\boxed{0}$	$\boxed{0}$	0	$\boxed{0}$	0	0	0	0
$\boxed{0}$	$\boxed{0}$	0	$\boxed{0}$	1	0	0	0
$\boxed{0}$	$\boxed{0}$	1	$\boxed{0}$	0	0	0	0
$\boxed{0}$	$\boxed{0}$	1	$\boxed{0}$	1	0	0	0
	$\vdots$		$\vdots$			$\vdots$	
$\boxed{1}$	0	$\boxed{0}$	$\boxed{1}$	0	1	0	1
$\boxed{1}$	0	$\boxed{0}$	$\boxed{1}$	1	1	0	1
$\boxed{1}$	1	$\boxed{0}$	$\boxed{1}$	0	1	0	1
$\boxed{1}$	1	$\boxed{0}$	$\boxed{1}$	1	1	0	1
	$\vdots$		$\vdots$			$\vdots$	

Notes: The full diagram would have $2^{2T+1} = 2^5 = 32$ rows corresponding to all possible realizations of the unobservables $U_i \equiv (Y_{i0}, U_{i1}(0), U_{i2}(0), U_{i1}(1), U_{i2}(1))$. Here, the rows shown are those corresponding to the potential outcomes that could generate $Y_i = (0, 0, 0)$ or $Y_i = (1, 0, 1)$ through the recursive relationship (1), i.e. the elements of $\mathcal{U}_{\text{oeq}}(0, 0, 0)$ and $\mathcal{U}_{\text{oeq}}(1, 0, 1)$ in (11). The observed outcome of Y_i is determined by the potential outcomes that are in boxes, but is unaffected by the other potential outcomes.

as a linear function of P:

$$\mathbb{P}[Y_i = y, X_i = x] = \sum_{u \in \mathcal{U}_{\text{oeq}}(y)} P(u, x), \tag{11}$$

where $\mathcal{U}_{\text{oeq}}(y)$ is the set of all $u \equiv (u_0, u_1(0), \ldots, u_T(0), u_1(1), \ldots, u_T(1)) \in \mathcal{U}$ for which $u_0 = y_0$ and $u_t(y_{t-1}) = y_t$ for all $t \geq 1$. Figure 1 illustrates (11) for $T = 2$.

Usually, a researcher is interested in a low-dimensional *target parameter*, that is, a low-dimensional function $\theta : \mathcal{P} \to \mathbb{R}^{d_\theta}$ of P. The researcher's primary object of interest is then the identified set for θ, denoted by $\Theta^\star \equiv \{\theta(P) : P \in \mathcal{P}^\star\}$. In the next section, I discuss some target parameters that provide useful measures of state dependence.

3.2 Target Parameters for Measuring State Dependence

A natural measure of state dependence is the proportion of agents that would have experienced a different outcome in period t had their outcome in period $t - 1$ been

different, i.e. the proportion of agents for which $U_{it}(0) \neq U_{it}(1)$. These agents are represented by the events $[U_{it}(0) = 0, U_{it}(1) = 1]$ and $[U_{it}(0) = 1, U_{it}(1) = 0]$. The proportion of the first group under P is denoted by

$$\mathrm{SD}_t^+(P) \equiv \mathbb{P}_P\left[U_{it}(0) = 0, U_{it}(1) = 1\right].$$

Agents in this first group can be said to experience positive state dependence, since an exogenous manipulation of their period $t-1$ outcome from 0 to 1 would result in a strictly positive increase in their period t outcome from 0 to 1. The measure of the second group under P is denoted by

$$\mathrm{SD}_t^-(P) \equiv \mathbb{P}_P\left[U_{it}(0) = 1, U_{it}(1) = 0\right].$$

This is the proportion of agents who can be said to experience negative state dependence. The total proportion of agents experiencing state dependence under P is

$$\mathrm{SD}_t(P) \equiv \mathbb{P}_P[U_{it}(0) \neq U_{it}(1)] = \mathrm{SD}_t^+(P) + \mathrm{SD}_t^-(P).$$

The average treatment effect of $Y_{i(t-1)}$ on Y_{it} is defined as

$$\mathrm{ATE}_t(P) \equiv \mathbb{E}_P[U_{it}(1) - U_{it}(0)],$$

where $\mathbb{E}_P$ denotes expectation taken with respect to P. This parameter is widely used to study state dependence, however it confounds two effects. To see this, notice that the relationship between ATE_t, SD_t^+, and SD_t^- is given by

$$\begin{aligned}
\mathrm{ATE}_t(P) &\equiv \mathbb{P}_P[U_{it}(1) = 1] - \mathbb{P}_P[U_{it}(0) = 1] \\
&= \left(\mathbb{P}_P\left[U_{it}(1) = 1, U_{it}(0) = 0\right] + \mathbb{P}_P[U_{it}(1) = 1, U_{it}(0) = 1]\right) \\
&\quad - \left(\mathbb{P}_P[U_{it}(1) = 0, U_{it}(0) = 1] + \mathbb{P}_P[U_{it}(1) = 1, U_{it}(0) = 1]\right) \\
&= \mathrm{SD}_t^+(P) - \mathrm{SD}_t^-(P). \qquad (12)
\end{aligned}$$

Thus, ATE_t is the proportion of the population that experiences positive state dependence at time t, less the proportion that experiences negative state dependence.

An implication of (12) is that it is possible for ATE_t to be small or zero even if there is substantial positive and negative state dependence. In such cases, ATE_t may be a misleading measure of state dependence. For example, suppose that Y_{it} denotes welfare status as in Chay et al. (2004) or Card and Hyslop (2005). Then SD_t^- represents the "at risk" proportion of the population that would receive welfare on period t as a

direct result of having not received it in the previous period, while SD_t^+ represents the proportion of the population that are in the "welfare trap." Both populations may be sizable, in which case ATE_t will not be a useful measure of state dependence. For this reason, I do not analyze ATE_t in this paper.

In many settings, it is interesting to consider modifying SD_t^+ and SD_t^- to be conditional on realizations of Y_{it}. For example, if Y_{it} is welfare status, a researcher may be interested in identifying positive state dependence among just the individuals currently receiving welfare, i.e. those with $Y_{it} = 1$. This parameter is given by

$$\text{SD}_t^+(P|1) \equiv \mathbb{P}_P \left[U_{it}(0) = 0, U_{it}(1) = 1 | Y_{it} = 1\right].$$

Alternatively, in the application to employment dynamics in Section 5, I consider positive state dependence among the unemployed, which is given by

$$\text{SD}_t^+(P|0) \equiv \mathbb{P}_P \left[U_{it}(0) = 0, U_{it}(1) = 1 | Y_{it} = 0\right].$$

These parameters are analogous to the treatment on the (un)treated parameters commonly considered in the analysis of the static potential outcome models (see e.g. Heckman and Vytlacil, 2007).

This type of conditioning can be extended a period further to define

$$\begin{aligned} &\text{SD}_t^+(P|00) \equiv \mathbb{P}_P \left[U_{it}(0) = 0, U_{it}(1) = 1 \middle| Y_{it} = 0, Y_{i(t-1)} = 0\right] \\ \text{and} \quad &\text{SD}_t^+(P|11) \equiv \mathbb{P}_P \left[U_{it}(0) = 0, U_{it}(1) = 1 \middle| Y_{it} = 1, Y_{i(t-1)} = 1\right], \end{aligned}$$

which quantify positive state dependence among agents whose state in the previous period is the same as in the current period. Like $\text{SD}_t^+(\cdot|0)$ and $\text{SD}_t^+(\cdot|1)$, these parameters have a treatment on the (un)treated interpretation. However, they can also be interpreted as the *proportion* of the observed persistence in outcomes that is due to state dependence.

To see this, consider the quantity $\mathbb{P}[Y_{it} = 0 | Y_{i(t-1)} = 0]$ as a measure of the observed persistence in state 0. For an observationally equivalent P, this quantity can be decomposed as

$$\begin{aligned} &\mathbb{P}[Y_{it} = 0 | Y_{i(t-1)} = 0] \\ &\quad = \mathbb{P}_P[Y_{it} = 0, U_{it}(1) = 0 | Y_{i(t-1)} = 0] + \mathbb{P}_P[Y_{it} = 0, U_{it}(1) = 1 | Y_{i(t-1)} = 0] \\ &\quad = \mathbb{P}_P[U_{it}(0) = 0, U_{it}(1) = 0 | Y_{i(t-1)} = 0] + \mathbb{P}_P[U_{it}(0) = 0, U_{it}(1) = 1 | Y_{i(t-1)} = 0]. \end{aligned}$$

The second term is the contribution to $\mathbb{P}[Y_{it} = 0 | Y_{i(t-1)} = 0]$ that is due to positive

state dependence. The size of this quantity as a proportion of the observed persistence is given by

$$\frac{\mathbb{P}_P[Y_{it}=0, U_{it}(1)=1|Y_{i(t-1)}=0]}{\mathbb{P}[Y_{it}=0|Y_{i(t-1)}=0]} = \mathbb{P}_P[U_{it}(1)=1|Y_{it}=0, Y_{i(t-1)}=0] = \text{SD}_t^+(P|00),$$

where the second equality follows because $[Y_{i(t-1)}=0, Y_{it}=0]$ implies $[U_{it}(0)=0]$. A similar argument shows that $\text{SD}_t^+(\cdot|11)$ can be interpreted as the proportion of the observed persistence in state 1 that is due to positive state dependence. These parameters constitute a natural rubric for measuring the role of state dependence in the persistence of observed outcomes.

3.3 Empirical-Evidence-Only Bounds

The data alone does not provide enough information to point identify SD_t^+ or SD_t^-. The reasons are the same as in a static potential outcomes model. First, an analyst never observes both $U_{it}(0)$ and $U_{it}(1)$, since only $Y_{it} = U_{it}(Y_{i(t-1)})$ is observed. Thus, quantities like SD_t^+ which concern the joint distribution of $(U_{it}(0), U_{it}(1))$ are inherently not point identified (see e.g. Heckman, Smith, and Clements, 1997). Second, even the marginal distributions of $U_{it}(0)$ and $U_{it}(1)$ will typically not be point identified due to the endogeneity of prior outcomes. That is, in general we expect that for observationally equivalent P,

$$\mathbb{P}[Y_{it}=1|Y_{i(t-1)}=1, X_i] = \mathbb{P}_P[U_{it}(1)=1|Y_{i(t-1)}=1, X_i] \neq \mathbb{P}_P[U_{it}(1)=1|X_i], \quad (13)$$

since $Y_{i(t-1)}=1$ depends on $(U_{i(t-1)}(0), U_{i(t-1)}(1))$, and $U_{it}(1)$ is likely dependent with $(U_{i(t-1)}(0), U_{i(t-1)}(1))$, even conditional on X_i, due to persistent latent heterogeneity.

While SD_t^+ and SD_t^- are not point identified, they are not completely unconstrained by the data. The next proposition provides sharp bounds on SD_t^+, SD_t^- and SD_t that use only the empirical evidence. All proofs are contained in Appendix D.

Proposition 1. *Suppose that $\mathcal{P}^\dagger = \mathcal{P}$. If $\theta = SD_t^+$, then*

$$\Theta^\star = \Big[0,\ \mathbb{P}[Y_{i(t-1)}=0, Y_{it}=0] + \mathbb{P}[Y_{i(t-1)}=1, Y_{it}=1]\Big]. \quad (14)$$

If $\theta = SD_t^-$, then

$$\Theta^\star = \Big[0,\ \mathbb{P}\left[Y_{i(t-1)}=0, Y_{it}=1\right] + \mathbb{P}\left[Y_{i(t-1)}=1, Y_{it}=0\right]\Big]. \quad (15)$$

If $\theta = SD_t$, then $\Theta^\star = [0,1]$.

The intuition behind the bounds in (14) is as follows. The target parameter is SD_t^+, which is the proportion of individuals with $U_{it}(0) = 0$ and $U_{it}(1) = 1$. Individuals with $Y_{i(t-1)} = 0$ and $Y_{it} = 1$ cannot have these potential outcomes, since they must have $U_{it}(0) = 1$. Similarly, individuals with $Y_{i(t-1)} = 1$ and $Y_{it} = 0$ have $U_{it}(1) = 0$, so they also do not contribute to SD_t^+. So, both of these observed groups can be removed from the calculation in (14).

The remaining observed groups are those with $Y_{i(t-1)} = Y_{it}$. Among individuals with $Y_{i(t-1)} = 0$ and $Y_{it} = 0$, there are those who have positive state dependence ($U_{it}(1) = 1$), and those who would have been in state 0 regardless ($U_{it}(1) = 0$). Similarly, the group with $Y_{i(t-1)} = 1$ and $Y_{it} = 1$ consists of individuals both with and without positive state dependence. The upper bound in (14) is obtained when all individuals in both observed groups exhibit positive state dependence, while the lower bound is obtained when none do. Analogous reasoning leads to the bounds in (15).

The lower bounds in (14)–(15) are always 0, regardless of the distribution of the data. Thus, the empirical evidence alone never enables a rejection of the hypothesis that there no positive or negative state dependence. The upper bound in (14) will be large when the observed outcomes have strong positive serial dependence. Similarly, the upper bound in (15) will be wide when the observed outcomes have strong negative serial dependence. Bounds on SD_t^+ will therefore tend to be wide when bounds on SD_t^- are narrow, and vice versa.

In fact, the third finding of Proposition 1 is that these upper bounds perfectly offset each other, so that the (sharp) identified set for state dependence of both sorts, SD_t, is always the entire logically possible interval $[0, 1]$. Thus, empirical evidence alone cannot discriminate between the hypothesis that there is no state dependence of any sort, with all persistence caused by latent heterogeneity, and the reverse, that there is no latent heterogeneity and all persistence in the data is due to state dependence. The existence or non-existence of state dependence can only be established by incorporating additional identifying assumptions, such as those discussed ahead.[13] In the next section, I provide a computational approach to constructing identified sets that can be used to flexibly combine multiple such assumptions.

3.4 Computing Identified Sets

Proposition 1 was proven using a standard two-step argument. First, one proposes bounds. Second, one shows that there are parameter values for which these bounds are obtained. This strategy provides analytic expressions, which can be useful both

[13] A similar point was made previously by Manski (2006).

for intuition and for statistical inference. However, the argument becomes increasingly complicated as the parameter space becomes increasingly complex.[14] Yet, the take-away from Proposition 1 was that we need to impose more assumptions in order to obtain interesting empirical conclusions. Since more assumptions typically make the parameter space more complex, this creates a problem.

One way to resolve this problem is to recognize that when θ is scalar-valued, its identified set can usually be determined by solving two optimization problems.[15]

Proposition 2. *Suppose that $\mathcal{P}^\dagger$ is closed and convex, and that θ is a continuous, scalar-valued function of P. Then, as long as $\mathcal{P}^\star$ is nonempty, $\Theta^\star = [\theta^\star_{lb}, \theta^\star_{ub}]$, where*

$$\theta^\star_{lb} \equiv \min_{P \in \mathcal{P}^\star} \theta(P) = \min_{\{P(u,x) \in [0,1] : u \in \mathcal{U}, x \in \mathcal{X}\}} \theta(P) \text{ s.t. } \rho(P) \geq 0, \text{ (10), and (11) } \forall y, x$$

$$\text{and } \theta^\star_{ub} \equiv \max_{P \in \mathcal{P}^\star} \theta(P) = \max_{\{P(u,x) \in [0,1] : u \in \mathcal{U}, x \in \mathcal{X}\}} \theta(P) \text{ s.t. } \rho(P) \geq 0, \text{ (10), and (11) } \forall y, x.$$

Proposition 2 suggests a computational approach to identification. The feasibility of this approach depends on how difficult it is to solve the problems that define $\theta^\star_{\text{lb}}$ and $\theta^\star_{\text{ub}}$. Observe that (11) places linear restrictions on $P = \{P(u,x) : u \in \mathcal{U}, x \in \mathcal{X}\}$. The requirement that $P \in \mathcal{P}$ also places linear restrictions on P, namely (10) and $1 \geq P(u,x) \geq 0$ for all $u \in \mathcal{U}$, $x \in \mathcal{X}$. Thus, if ρ and θ are linear, then the two optimization problems in Proposition 2 are linear programs. This means that Proposition 2 can be applied even in high dimensions as long as we limit attention to parameters and assumptions that can be expressed as linear functions of P.

Linearity turns out to be not very restrictive in the sense that it still permits a wide range of interesting parameters and assumptions. In Appendix E, I show that each of the target parameters discussed in Section 3.2 is a linear function of P. When considering additional prior assumptions in Section 4, I will restrict attention to those that are linear due to computational considerations. This is not essential to Proposition 2, which applies more generally, but it is important for practical implementation.

[14] For example, compare the analysis in Okumura and Usui (2014) to that in Manski and Pepper (2000) and Manski (1994), or the analysis of Mourifié (2015) to that of Shaikh and Vytlacil (2011).

[15] This general point about partial identification analysis has been appreciated (sometimes implicitly) by many previous authors, including Honoré and Tamer (2006), Manski (2007), Molinari (2008), Chiburis (2010), Kitamura and Stoye (2013), Manski (2014), Freyberger and Horowitz (2015) and Laffers (2013, 2018). In particular, Laffers (2013, 2018) uses a similar computational strategy as in this paper, but for a static potential outcomes model; see also the subsequent work by Demuynck (2015). The benefits in the static setting are smaller than in the dynamic case considered here, since a large number of analytic partial identification results already exist for static potential outcomes models. The representation of bounds in terms of linear programming is not new to this paper, and dates back to at least Balke and Pearl (1994, 1997) for similar problems in causal inference, or to Hansen, Heaton, and Luttmer (1995) for different problems in finance.

4 Identifying Assumptions for the DPO Model

The empirical-evidence-only identified sets derived in Proposition 1 are wide. In this section, I propose identifying assumptions that can be added to the DPO model to narrow these bounds. These assumptions are implemented by including restrictions in the ρ function of Section 3 and then applying Proposition 2. Any subset of these assumptions can be combined by simply adding or removing the appropriate restrictions. All assumptions could be modified to be conditional on covariates (X_i). I keep this implicit for the sake of notation, but indicate situations in which conditioning may be important.[16] Along the way, I motivate and interpret the assumptions in the context of the DC and DBR models discussed in Section 2.

4.1 Stationarity

Stationarity assumptions are ubiquitous in panel data models. Indeed, combining empirical evidence from different time periods *requires* an assumption that the past shares at least some features in common with the future. In the DPO model, one stationarity assumption is that the joint distribution of $(U_{it}(0), U_{it}(1))$ is invariant across $t \geq 1$. A stronger form of stationarity uses multiple time periods, e.g. that the distribution of $(U_{i(t-1)}(0), U_{it}(0), U_{i(t-1)}(1), U_{it}(1))$ does not vary across $t \geq 2$. The following is a general version.

Assumption ST: *Let m be a non-negative integer chosen by the researcher and define $U_{i(t-m:t)}(0) \equiv (U_{i(t-m)}(0), \ldots, U_{it}(0))$ and $U_{i(t-m:t)}(1) \equiv (U_{i(t-m)}(1), \ldots, U_{it}(1))$ for $t \geq m+1$. For any $u \equiv (u_m(0), u_m(1)) \in \{0,1\}^{2(m+1)}$ define*

$$\Sigma^u_{t,m}(P) \equiv \mathbb{P}_P[U_{i(t-m:t)}(0) = u_m(0), U_{i(t-m:t)}(1) = u_m(1)].$$

Then for any $P \in \mathcal{P}^\dagger$, every $u \in \{0,1\}^{2(m+1)}$, and every $t, t' \geq m+1$,

$$\Sigma^u_{t,m}(P) = \Sigma^u_{t',m}(P). \tag{16}$$

Distributions of potential outcomes that satisfy Assumption ST do not need to generate distributions of observed outcomes that are stationary. To see this, first

[16] Note that including rich covariate specifications quickly increases the dimension of the problems in Proposition 2. Three dimension reduction strategies are discussed in Appendix F.

observe that for any P,

$$\begin{aligned}\mathbb{P}_P[Y_{it}=0] &= \mathbb{P}_P[U_{it}(0)=U_{it}(1), Y_{it}=0] + \mathbb{P}_P[U_{it}(0)\neq U_{it}(1), Y_{it}=0] \\ &= \mathbb{P}_P[U_{it}(0)=0, U_{it}(1)=0] + \mathbb{P}_P[U_{it}(0)\neq U_{it}(1), Y_{it}=0], \qquad (17)\end{aligned}$$

where the second equality follows because $\mathbb{P}_P[U_{it}(0)=U_{it}(1)=1, Y_{it}=0]=0$ by (1). Thus, if P satisfies Assumption ST, then from (17) there is marginal stationarity in the observed outcomes, i.e. $\mathbb{P}_P[Y_{it}=0]=\mathbb{P}_P[Y_{i(t-1)}=0]$, if and only if

$$\mathbb{P}_P[U_{it}(0)\neq U_{it}(1), Y_{it}=0] = \mathbb{P}_P[U_{i(t-1)}(0)\neq U_{i(t-1)}(1), Y_{i(t-1)}=0]. \qquad (18)$$

This restriction is not generally implied by Assumption ST. For (18) to hold would require a condition about the serial dependence of the potential outcomes at all previous lags, whereas Assumption ST does not restrict this dependence.

However, one case in which (18) does need to be true is when there is no state dependence, so that $\text{SD}_{t'}(P) \equiv \mathbb{P}_P[U_{t'}(0)\neq U_{t'}(1)]=0$ for $t'=t-1,t$. An important implication is that if Assumption ST is maintained, and if the distribution of observed outcomes is *not* stationary, then it is possible to rule out the hypothesis that the identified set for SD_t contains 0 or other small values. Intuitively, in order for a DPO model that satisfies Assumption ST to generate a stationary distribution of observed outcomes, it must be the case that there is also no state dependence. To the extent that the observed outcome sequence is in fact nonstationary, one can therefore rule out the hypothesis that there is no state dependence.[17]

The potential outcomes implied by the DC model through (7) depend on $\Delta\mathring{\nu}(S_{it}(y))$. Whether $\Delta\mathring{\nu}(S_{it}(y))$ is stationary depends on the form of $\Delta\mathring{\nu}$ and the composition of the counterfactual state variables $S_{it}(y) \equiv (y, d^\star(S_{i(t-1)}\|y), Z_{it})$. Certainly, any time-invariant state variable is trivially stationary. It is also standard in empirical implementations of DC models to assume that the unobservable state variables are stationary. However, it may be undesirable to assume that time-varying components of the observed state variables are stationary, since this could be directly rejected by the data. In these cases, one can impose a version of Assumption ST that conditions on these variables.

Another conceptual issue highlighted by the DC model is the distinction between finite and infinite horizons. A finite horizon assumption can be viewed in terms of an infinite horizon problem by defining the flow utility to be 0 after the finite horizon

[17] A similar observation was used by Heckman (1981b, pg. 159) to establish point identification of β_0 in the parametric DBR model (8).

has elapsed. This can be captured by an "age" state variable in Z_{it} that records the agent's location in their finite horizon. The value function—and therefore the potential outcomes $U_{it}(y)$—cannot be unconditionally stationary with a finite horizon, since it becomes identically 0 after a given time period. If a finite horizon is empirically important, then Assumption ST can be modified to be conditional on age.

4.2 Diminishing Serial Correlation

Persistent heterogeneity may cause potential outcomes to be positively serially correlated. If transitory heterogeneity is also present, then it may be natural to assume that this serial correlation is strongest between potential outcomes in adjacent periods and diminishes (or does not increase) as the distance between any two periods increases. This is the content of the following assumption.

Assumption DSC: *For every $P \in \mathcal{P}^\dagger$, and each $y \in \{0,1\}$, $\text{Corr}_P(U_{it}(y), U_{i(t+t')}(y))$ is decreasing in $|t'|$ for $t' \in \{1-t, \ldots, T-t\}$.*

Assumption DSC places a nonlinear restriction on P. However, if Assumption ST holds (with any $m \geq 0$) then Assumption DSC becomes a linear restriction, equivalent to the statement that $\mathbb{P}_P[U_{it}(y) = 1, U_{i(t+t')}(y) = 1]$ is decreasing in $|t'|$ for $t' \in \{1-t, \ldots, T-t\}$.[18] Due to this computational consideration, I will only consider imposing Assumption DSC when it is combined with Assumption ST.

The following proposition provides sufficient conditions for Assumption DSC when potential outcomes are generated by a DC model through (7).

Proposition 3. *Suppose that Assumption ST holds with any $m \geq 0$. Then Assumption DSC is satisfied if every $P \in \mathcal{P}^\dagger$ is consistent with the following conditions for each fixed $y \in \{0,1\}$: (i) $U_{it}(y)$ is determined by (7); (ii) There is a function φ, time-invariant random variables $\bar{S}_i$, and scalar time-varying random variables, $\tilde{S}_{it}$, such that $\Delta\mathring{\nu}(S_{it}(y)) = \varphi(\bar{S}_i, \tilde{S}_{it})$, where $\varphi(\bar{s}, \cdot)$ is weakly increasing and right-continuous for each $\bar{s}$; and (iii) $\{\tilde{S}_{it}\}_{t=1}^T | \bar{S}_i$ is a first-order Markov chain with $\mathbb{P}[\tilde{S}_{it} \leq \tilde{s}_t | \tilde{S}_{i(t-1)} = \tilde{s}_{t-1}, \bar{S}_i = \bar{s}]$ weakly decreasing in $\tilde{s}_{t-1}$ for all $\tilde{s}_t$ and $\bar{s}$.*

Proposition 3 applies to the DBR model, (9), by taking $\bar{S}_i = A_i$, $\tilde{S}_{it} = X_{it}'\beta_1 + V_{it}$, and $\varphi(\bar{s}, \tilde{s}) \equiv \beta_0 y + \bar{s} + \tilde{s}$. The sufficient condition for Assumption DSC is that $X_{it}'\beta_1 + V_{it}$ is a first-order Markov chain with a stochastically increasing transition distribution. If

[18] This statement is justified in Appendix E. If Assumption ST does not hold, then the statement that $\mathbb{P}_P[U_{it}(y) = 1, U_{i(t+t')}(y) = 1]$ is decreasing in $|t'|$ is equivalent to the statement that $(U_{it}(y), U_{i(t+t')}(y))$ is decreasing in the upper orthant order with respect to $|t'|$, see e.g. Shaked and Shanthikumar (2007, Section 6.G). The upper orthant order does not necessarily have a clear interpretation as a positive dependence concept, so imposing this condition directly does not seem attractive.

there are no covariates ($\beta_1 = 0$), and $(V_{it}, V_{i(t+1)})$ are jointly normal (conditional on A_i), then the stochastic increasing assumption is equivalent to the correlation between V_{it} and $V_{i(t+1)}$ (given A_i) being non-negative.

4.3 Monotone Treatment Selection

For a static model, Manski and Pepper (2000) considered the identifying content of assuming that potential outcomes are greater for agents who select into treatment than for those who do not. This monotone treatment selection (MTS) condition captures the idea that a researcher may be willing to make a prior assumption on the direction of bias that would arise from a simple treatment–control contrast. The following is a similar assumption for the DPO model.

Assumption MTS. *Every $P \in \mathcal{P}^\dagger$ satisfies*

$$\mathbb{P}_P[U_{it}(y) = 1 | Y_{i(t-1)} = 1, Y_{i(t-2)} = \tilde{y}] \geq \mathbb{P}_P[U_{it}(y) = 1 | Y_{i(t-1)} = 0, Y_{i(t-2)} = \tilde{y}] \quad (19)$$

for $y = 0, 1$, $\tilde{y} = 0, 1$ and all $t \geq 2$ such that $\mathbb{P}[Y_{i(t-1)} = 1 | Y_{i(t-2)} = \tilde{y}] \in (0, 1)$.

Assumption MTS says that those with $Y_{i(t-1)} = 1$ would be more likely to have $Y_{it} = 1$ than those with $Y_{i(t-1)} = 0$, even if their outcomes in period $t-1$ were exogenously manipulated from 0 to 1 or vice versa. Stated differently, the assumption is that agents with $Y_{i(t-1)} = 1$ have a higher latent propensity to be in state 1 in period t than agents with $Y_{i(t-1)} = 0$. The additional conditioning on $Y_{i(t-2)} = y_{t-2}$ in these statements ensures that the outcome in year $t-1$ is comparable. That is, since the event $[Y_{i(t-1)} = y_{t-1}, Y_{i(t-2)} = y_{t-2}]$ is equivalent to the event $[U_{i(t-1)}(y_{t-2}) = y_{t-1}, Y_{i(t-2)} = y_{t-2}]$, conditioning on $Y_{i(t-2)} = y_{t-2}$ ensures that the conditioning events on the left and right sides of (19) are expressed in terms of the same potential outcome $U_{i(t-1)}(y_{t-2})$. A stronger form of Assumption MTS extends this conditioning back to period $t - q$.

Assumption MTS (generalization). *Let $q \geq 2$ be an integer chosen by the analyst. For each t, if $q < t$ then let $Y_{i(t-q):(t-2)} \equiv (Y_{i(t-q)}, \ldots, Y_{i(t-2)})$; otherwise let $Y_{i(t-q):(t-2)} \equiv (Y_0, \ldots, Y_{t-2})$. Every $P \in \mathcal{P}^\dagger$ satisfies*

$$\begin{aligned} &\mathbb{P}_P[U_{it}(y) = 1 | Y_{i(t-1)} = 1, Y_{i(t-q):(t-2)} = y_{past}] \\ &\quad \geq \mathbb{P}_P[U_{it}(y) = 1 | Y_{i(t-1)} = 0, Y_{i(t-q):(t-2)} = y_{past}] \end{aligned} \quad (20)$$

for $y = 0, 1$, $y_{past} \in \{0, 1\}^{\min\{q,t\}-1}$ and all $t \geq 2$ such that $\mathbb{P}[Y_{i(t-1)} = 1 | Y_{i(t-q):(t-2)} = y_{past}] \in (0, 1)$.

When potential outcomes are generated by the DC model (7), a sufficient condition for Assumption MTS is that $\Delta\mathring{\nu}(S_{it}(y))$ and $\Delta\mathring{\nu}(S_{it}(\tilde{y}))$ exhibit local positive quadrant dependence for each $(y, \tilde{y})$.

Proposition 4. *Assumption MTS is satisfied with $q = 2$ if every $P \in \mathcal{P}^{\dagger}$ is consistent with the following conditions for each $(y, \tilde{y}) \in \{0,1\}^2$: (i) $U_{it}(y)$ is determined by (7); (ii) $\Delta\mathring{\nu}(S_{it}(y))$ and $\Delta\mathring{\nu}(S_{it}(\tilde{y}))$ are positive quadrant dependent at $(0,0)$, conditional on $Y_{i(t-2)} = \tilde{y}$, i.e.*

$$\begin{aligned}&\mathbb{P}\left[\Delta\mathring{\nu}(S_{it}(y)) \geq 0,\ \Delta\mathring{\nu}(S_{i(t-1)}(\tilde{y})) \geq 0 \,\middle|\, Y_{i(t-2)} = \tilde{y}\right] \\ &\quad \geq \mathbb{P}\left[\Delta\mathring{\nu}(S_{it}(y)) \geq 0 \,\middle|\, Y_{i(t-2)} = \tilde{y}\right] \mathbb{P}\left[\Delta\mathring{\nu}(S_{i(t-1)}(\tilde{y})) \geq 0 \,\middle|\, Y_{i(t-2)} = \tilde{y}\right]. \qquad (21)\end{aligned}$$

Condition (ii) depends on the structure of $\Delta\mathring{\nu}$, as well as the composition and relationships among the state variables, $S_{it}(y)$. In Section 5, I discuss a job search model in which this condition can be made more primitive. The effective requirement is that there is positive dependence among the determinants of being in state 1.

In the DBR model (8), Assumption MTS is satisfied if $(X_{it}'\beta_1 + A_i + V_{it})$ and $(X_{i(t-1)}'\beta_1 + A_i + V_{i(t-1)})$ are locally positive quadrant dependent, conditional on $Y_{i(t-2)}$. If there are no covariates ($\beta_1 = 0$), and if $(V_{it}, V_{i(t-1)})$ and A_i are independent and normally distributed conditional on $Y_{i(t-2)}$, then this will be the case if V_{it} and $V_{i(t-1)}$ are weakly positively correlated, or even negatively correlated, so long as the magnitude of the covariance between V_{it} and $V_{i(t-1)}$ is smaller than the variance of A_i.

4.4 Fixed Effects

Another way to introduce a permanent-transitory distinction in unobserved heterogeneity is the following assumption, introduced by Chernozhukov, Fernández-Val, Hahn, and Newey (2013). Those authors describe it as "time is randomly assigned" or "time is an instrument" (TIV).

Assumption TIV. *Let $U_{it} \equiv (U_{it}(0), U_{it}(1))$. For every $P \in \mathcal{P}^{\dagger}$, there exists a random variable A_i such that if U_i is distributed according to P and Y_i is generated by (1), then*

$$\mathbb{P}[U_{it} = u | Y_{i(t-1)}, \ldots, Y_{i1}, Y_{i0}, A_i] = \mathbb{P}[U_{i1} = u | Y_{i0}, A_i] \quad \textit{(almost surely)}$$

for all $u \in \{0,1\}^2$ and all $t \geq 2$.

Assumption TIV implies that all persistent unobservable heterogeneity is captured by the time-invariant latent random variable A_i, which can be interpreted as a generaliza-

tion of a fixed effect. After accounting for A_i and the initial state Y_{i0}, current potential outcomes are required to be independent of past realized outcomes.

Although A_i has no meaning itself within the DPO model, Assumption TIV implies many restrictions on the distribution of potential outcomes.

Proposition 5. *Let $Y_{i(0:t)} \equiv (Y_{i0}, Y_{i1}, \ldots, Y_{it})$. If Assumption TIV holds then for every $P \in \mathcal{P}^\dagger$, and every $t' > t \geq 1$,*

$$\mathbb{P}_P[U_{it'} = u, Y_{i(0:t-1)} = y] = \mathbb{P}_P[U_{it} = u, Y_{i(0:t-1)} = y] \tag{22}$$

for all $u \in \{0,1\}^2$ and $y \in \{0,1\}^t$. As a consequence, Assumption TIV implies Assumption ST with $m = 0$.

I have been unable to determine whether the converse of Proposition 5 is also true.[19] If it is not, then imposing (22) might yield a non-sharp identified set. In the application in Section 5, I test the implications of TIV in Proposition 5 and overwhelmingly reject them.

When potential outcomes are generated by the DC model, the following conditions are sufficient for Assumption TIV and therefore (22).

Proposition 6. *Assumption TIV is satisfied if every $P \in \mathcal{P}^\dagger$ is consistent with the following conditions: (i) $U_{it}(y)$ is determined by (7); and (ii) The stochastic components of $(S_{it}(0), S_{it}(1))$ can be split into time-invariant components, $\bar{S}_i$, and time-varying components, $\tilde{S}_{it}$, in such a way that the distribution of $\tilde{S}_{it}|\tilde{S}_{i(1:t-1)}, \bar{S}_i, Y_{i0}$ is the same as that of $\tilde{S}_{i1}|\bar{S}_i, Y_{i0}$ for every $t \geq 2$.*

4.5 Monotone Instrumental Variables

Instrumental variables (IV) can be used by assuming that an observed state variable (the instrument) is independent of potential outcomes. The monotone instrumental variable (MIV) assumption introduced by Manski and Pepper (2000, 2009) weakens this assumption to only restrict the sign of the relationship between the potential outcomes and instrument. The following is one way to adapt the MIV assumption to the DPO model.

Assumption MIV: *Let X_{it}^0 and X_{it}^1 be subvectors of X_i, where X_{it}^1 takes values in a partially ordered set. Every $P \in \mathcal{P}^\dagger$ is such that $\mathbb{P}_P[U_{it}(y) = 1 | X_{it}^0 = x^0, X_{it}^1 = x^1]$ is weakly increasing (or decreasing) in x^1 for every x^0, each $y = 0, 1$, and every $t \geq 1$.*

[19] That is, if P satisfies the condition in Proposition 5, does there exist a random variable A_i such that P satisfies the condition in Assumption TIV?

Assumption MIV can be strengthened to a full IV assumption by imposing both directions of weak monotonicity.

The next proposition provides sufficient conditions for Assumption MIV when potential outcomes are generated by the DC model.

Proposition 7. *Assumption MIV is satisfied if every $P \in \mathcal{P}^\dagger$ is consistent with the following conditions for each fixed $y \in \{0,1\}$: (i) $U_{it}(y)$ is determined by (7); (ii) $S_{it}(y)$ can be partitioned as $S_{it}(y) = (X_{it}^0, X_{it}^1, V_{it})$, where X_{it}^0 and X_{it}^1 are observed to the researcher, V_{it} is unobserved, and X_{it}^1 takes values in a partially ordered set; (iii) X_{it}^1 is independent of V_{it}, conditional on X_{it}^0; and (iv) $\Delta\mathring{\nu}(S_{it}(y))$ can be written as $\Delta\mathring{\nu}(S_{it}(y)) = \varphi(X_{it}^0, X_{it}^1, V_{it})$ for a function φ that is increasing (or decreasing) in X_{it}^1 for all fixed X_{it}^0 and V_{it}.*

Common implementations of the DBR model (8) assume that A_i is independent of X_{it}, and V_{it} is independent of (X_{it}, A_i) with (known) distribution Φ. In this case,

$$\mathbb{P}[U_{it}(y) = 1 | X_{it} = x] = \mathbb{E}\left[\Phi(\beta_0 y + x'\beta_1 + A_i)\right]. \tag{23}$$

Thus, the DBR model implies that $\mathbb{P}[U_{it}(y) = 1 | X_{it} = x]$ is monotone increasing in a component of x if the sign of the corresponding component of β_1 is positive. Assumption MIV amounts to placing a sign restriction on a component of β_1. Imposing Assumption MIV in both directions—i.e. assuming that a particular component of β_1 is 0—corresponds to an exclusion restriction. Exclusion restrictions like these are often imposed in applications of DBR models by not including various leads and lags of the time-varying observables.

4.6 Monotone Treatment Response

In some applications, it may make sense to assume that state dependence is either positive or negative. This can be viewed as a dynamic version of Manski's (1997) monotone treatment response (MTR) assumption.

Assumption MTR: *Every $P \in \mathcal{P}^\dagger$ satisfies $\mathbb{P}_P[U_{it}(1) \geq U_{it}(0)] = 1$ for all t.*

If potential outcomes are determined by the DC model through (7), a sufficient condition for Assumption MTR is that $\Delta\mathring{\nu}(S_{it}(1)) \geq \Delta\mathring{\nu}(S_{it}(0))$ almost surely. Either this condition or its opposite is always satisfied in the special case of the DBR model (8)–(9), since the parameter β_0 on lagged outcomes is treated as deterministic.

A weaker version of Assumption MTR only imposes monotonicity "on average."

Assumption MATR: *Every $P \in \mathcal{P}^\dagger$ satisfies $\mathbb{P}_P[U_{it}(1) = 1] \geq \mathbb{P}_P[U_{it}(0) = 1]$ for all t.*

Assumption MATR is equivalent to the assumption that $\text{ATE}_t(P)$ is positive. In light of (12), this is equivalent to the assumption that there is more positive state dependence than there is negative state dependence. For the DPO model (7), a sufficient and necessary condition for Assumption MATR is that $\Delta\mathring{\nu}(S_{it}(1))$ is more likely than $\Delta\mathring{\nu}(S_{it}(0))$ to be greater than 0.

5 The Unemployment Dynamics of High School Educated Men

5.1 Background and Motivation

The increase of long-term unemployment in the wake of the Great Recession has rejuvenated interest on state dependence in unemployment. There is an extensive literature that estimates parametric DBR models with European employment data, e.g. Narendranathan and Elias (1993), Mühleisen and Zimmermann (1994), Arulampalam, Booth, and Taylor (2000), and Tumino (2015). This research typically finds substantial evidence of state dependence in unemployment. On the other hand, the comparably few studies that used similar methods with U.S. data, such as Ellwood (1982) and Corcoran and Hill (1985) find little or no evidence of state dependence.

A recent line of field experiments (Oberholzer-Gee, 2008; Kroft, Lange, and Notowidigdo, 2013; Ghayad, 2013; Eriksson and Rooth, 2014; Farber, Silverman, and von Wachter, 2016, 2017; Nunley, Pugh, Romero, and Seals, 2016; Farber, Herbst, Silverman, and von Wachter, 2018) have provided convincingly–identified nonparametric estimates of the causal effect of employment gaps in fictitious resumes on the callback rates of prospective employers. The evidence from this literature has been mixed, with some studies finding no evidence of short-term state dependence (Eriksson and Rooth, 2014; Nunley et al., 2016; Farber et al., 2017) and others finding negative effects (Kroft et al., 2013; Ghayad, 2013). Farber et al. (2018) examine potential explanations for these differences, but conclude that the contrast in results remains a puzzle.

In this section, I use the DPO model to take a different look at this topic. The analysis uses observational data, and so avoids a key criticism of the experimental literature that callbacks may have a limited relationship to actual employment outcomes (Jarosch and Pilossoph, forthcoming). This benefit to external validity comes at the cost of imposing assumptions that are less credible than random assignment in a controlled experiment. On the other hand, these assumptions are nonparametric, so they may be an attractive alternative to researchers who are concerned about the impact of specific functional forms. The cost of remaining nonparametric is the loss of point identification, but the results ahead show that one can nevertheless still obtain informative estimates.

Table 1: Descriptive Statistics on Unemployment Dynamics in the SIPP

	time period t						
	0	1	2	3	4	5	6
$\mathbb{P}[Y_{it} = 1]$	.921	.936	.945	.931	.945	.949	.942
	(.005)	(.004)	(.004)	(.004)	(.004)	(.004)	(.004)
$\mathbb{P}[Y_{it} \neq Y_{i(t-1)}]$	_	.067	.056	.055	.050	.043	.048
		(.004)	(.004)	(.004)	(.004)	(.003)	(.004)
$\mathbb{P}[Y_{it} = 0 \mid Y_{i(t-1)} = 0]$	_	.483	.493	.632	.534	.574	.594
		(.030)	(.034)	(.035)	(.032)	(.036)	(.037)
$\mathbb{P}[Y_{it} = 1 \mid Y_{i(t-1)} = 1]$	_	.972	.975	.964	.981	.979	.971
		(.003)	(.003)	(.003)	(.002)	(.002)	(.003)
naive ATE	_	.455	.468	.595	.515	.554	.565
		(.030)	(.034)	(.035)	(.032)	(.036)	(.037)

	percentage of agents with ...							
	0	1	2	3	4	5	6	7
periods of unemployment	82.30	7.83	3.29	2.36	1.75	1.11	0.47	0.90
unemployment spells	82.30	13.36	3.61	0.73	0.00	–	–	–
transitions	83.20	6.78	6.43	2.21	1.25	0.12	0.00	–

Notes: Standard errors are given in parentheses. A transition is defined as the event $[Y_{it} \neq Y_{i(t-1)}]$. The "naive ATE" is defined as $\mathbb{P}[Y_{it} = 1 \mid Y_{i(t-1)} = 1] - \mathbb{P}[Y_{it} = 1 \mid Y_{i(t-1)} = 0]$. The sample size is 3,435.

5.2 Data

The dataset is an extract of the 2008 Survey of Income and Program Participation (SIPP), which is a nationally-representative longitudinal survey covering the period of September 2008 to December 2013. Individuals in the SIPP are surveyed in four month waves and questioned retrospectively on their employment status at different points over the previous four months. In order to mitigate seam bias, I follow a conservative strategy of using only observations corresponding to the last month of the retrospective four month interview period (Grogger, 2004; Ham and Shore-Sheppard, 2005). Also, I limit my sample to the period between January 2011 and April 2013, so as to avoid the more turbulent times surrounding the Great Recession. This leaves seven ($T = 6$) periods that are four months long each.[20]

I used the following sample selection rules when constructing the extract. I restricted attention to the subpopulation of working age men who were between 18 and

[20] I did not include the final two survey waves because they have unusually high attrition rates.

55 years of age at the time of the initial survey. I kept only men who reported either being employed or actively searching for work during all periods, so that $Y_{it} = 1$ denotes employment and $Y_{it} = 0$ denotes unemployment.[21] I also limit the focus to men who had a high school education or the equivalent, but no college degree, and who were not enrolled in school at any point during the sample. I dropped men who reported having a work-preventing disability or serving in the military at any point in the sample. Also, I removed observations that were heavily imputed (referred to as "type z" in the SIPP) or had other indications of irregularity.

After balancing the panel, the cross section consists of 3,435 men. Table 1 reports some summary statistics on the employment dynamics of these men. The overall unemployment rate ranges from roughly 5–8% over the course of the sample and employment is highly persistent with roughly 97% of men employed in one period remaining employed in the next. Unemployment is less persistent, with the probability of an unemployed man remaining unemployed ranging from approximately 50% early in the sample to around 60% in the later periods. Fewer than 1% of men remained unemployed in every period, but roughly 18% experience at least one spell of unemployment. A naive estimate of the average treatment effect would be between .455 and .595 depending on the period considered. This estimate probably overstates the role of positive state dependence if there is a permanent source of latent heterogeneity that positively affects the propensity to be employed.

5.3 A Model of Job Search with Endogenous Effort

In this section, I develop a DC model of job search to help motivate and interpret the assumptions used in the empirical analysis. The model features on-the-job search and endogenous search effort as in Christensen, Lentz, Mortensen, Neumann, and Werwatz (2005) and Faberman, Mueller, Şahin, and Topa (2017). The model is nonparametric with respect to the functional form of search effort costs, unobserved heterogeneity, and the distribution of wage offers. This generality is possible because the only way the job search model will be used is to motivate nonparametric assumptions for the DPO model.

Worker i begins period t having either been employed or unemployed in the previous period ($Y_{i(t-1)} = 1$ or 0, respectively), and having exerted $E_{i(t-1)}$ units of search effort in the previous period. The worker receives a wage offer, $\omega(Y_{i(t-1)}, E_{i(t-1)}, A_i, V_{it})$, which depends on their work and effort choices in the previous period, a permanent

[21] Following Chetty (2008), I classify a worker as employed if they report having a job the entire interview month and not being on layoff.

(time-invariant) source of heterogeneity, A_i, and a time-varying wage shock, V_{it}.[22] After observing the offer, the worker decides to either accept it and work in period t ($Y_{it} = 1$) or to remain unemployed ($Y_{it} = 0$).[23] In either case, they also decide on a level of search effort, E_{it}, to exert in period t as an investment to getting a better offer in period $t + 1$. The worker solves this problem with an infinite horizon ($\bar{T} = +\infty$).

The worker's flow utility from work decision y' and effort choice e' is given by

$$\mu\left(y', e', Y_{i(t-1)}, E_{i(t-1)}, A_i, V_{it}\right) = y'\omega\left(Y_{i(t-1)}, E_{i(t-1)}, A_i, V_{it}\right) - \kappa\left(y', e', A_i\right),$$

where $\kappa(y', e', A_i)$ is the cost of exerting e' units of search effort when making employment choice y', and A_i allows for heterogeneity in these costs across workers. Allowing both κ and ω to depend on employment status means that both the costs and efficacy of searching can vary for employed and unemployed workers, as in Faberman et al. (2017). These are the two potential sources of state dependence in the model.

The distribution of the wage shock, V_{it}, is assumed to be first-order Markov, conditional on permanent heterogeneity.

Assumption MC: *$\{V_{it}\}_{t=1}^{T}$ is first-order Markov, conditional on A_i.*

The worker is assumed to know the conditional distribution of V_{it}, so that under Assumption MC the Bellman equation is given by

$$\begin{aligned}
&\nu\left(Y_{i(t-1)}, E_{i(t-1)}, V_{it}, A_i\right) \\
&\quad = \max_{(y',e')\in\{0,1\}\times\mathcal{E}} \left\{ \begin{array}{c} y'\omega\left(Y_{i(t-1)}, E_{i(t-1)}, A_i, V_{it}\right) - \kappa\left(y', e', A_i\right) \\ + \delta\mathbb{E}\left[\nu\left(y', e', V_{i(t+1)}, A_i\right) \mid V_{it}, A_i\right] \end{array} \right\},
\end{aligned} \tag{24}$$

Thus, the state variables at time t are $S_{it} \equiv (Y_{i(t-1)}, E_{i(t-1)}, V_{it}, A_i)$.

This dynamic job search model generates a DPO model through (7). Assumptions about the job search model imply assumptions on the generated DPO model. Consider the following assumptions on the wage shocks.

Assumption W(a): *The distribution of $(V_{i(t-m')}, \ldots, V_{it})|A_i = a$ does not depend on t for any a, where m' is some non-negative integer.*

Assumption W(b): *V_{it} and $V_{i(t+1)}$ are independent, conditional on A_i, for all t.*

Assumption W(a) is that the wage offer distribution is stationary, which is a common assumption in empirical and theoretical analyses of labor search models, as well as

[22] Not receiving an offer (or being laid off) corresponds to receiving an offer of $-\infty$.
[23] Past offers cannot be recalled.

empirical implementations of dynamic discrete choice models more generally.[24] The assumption still allows the distribution of wage shocks to vary for workers with different time-invariant characteristics. Assumption W(b) is that the wage shocks are serially independent, conditional on these characteristics. This is also a common assumption in empirical implementations of dynamic discrete choice models.[25]

The next proposition shows that Assumptions W(a) and W(b) imply that the generated DPO model satisfies Assumptions ST and DSC, respectively. If both assumptions hold, and the time-invariant characteristics include the initial employment outcome, Y_{i0}, then the generated DPO model also satisfies Assumption TIV.

Proposition 8. *Suppose that every $P \in \mathcal{P}^{\dagger}$ is consistent with $U_{it}(y)$ being generated through* (7) *by the DC model described in this section. Suppose that Assumption MC is also satisfied.*

(i) *If Assumption W(a) is satisfied, then Assumption ST is satisfied with $m = m' - 1$.*

(ii) *If Assumption W(b) is satisfied, then Assumption DSC is satisfied.*

(iii) *If Assumptions W(a) and W(b) are satisfied, then Assumption ST is satisfied with $m = m'$.*

(iv) *Suppose that $A_i \equiv (\bar{A}_i, Y_{i0})$ contains the worker's employment choice in period* 0. *Then Assumption TIV is satisfied if Assumptions W(a) and Assumption W(b) are satisfied.*

Requiring Y_{i0} to be included as part of A_i in Proposition 8 (iv) is difficult to motivate, since in the current application period 0 simply reflects the first period of observing data, and not some initial period for the workers. The empirical results in the next section strongly reject the hypothesis that the model is correctly specified under Assumption TIV.

The next proposition provides sufficient conditions for Assumption MTS.

Proposition 9. *Suppose that every $P \in \mathcal{P}^{\dagger}$ is consistent with $U_{it}(y)$ being generated through* (7) *by the DC model described in this section. Suppose that Assumptions MC and W(b) are satisfied, and that the following additional conditions hold: (i) A_i takes values in a partially ordered set, and it includes the initial conditions $(Y_{i\underline{T}}, E_{i\underline{T}})$ at the beginning of the choice problem; (ii) A_i and V_{it} are independent; (iii) There is no*

[24] As Keane et al. (2011, pg. 371) write when discussing stationarity, "Most DCDP [discrete choice dynamic programming] models in the literature which solve the full dynamic programming problem implicitly make such an assumption as well, though it is not dictated by the method."

[25] However, see Norets (2009) and Connault (2016) for approaches to incorporating serially correlated unobservables.

search effort decision, so that $\kappa(y', e', a) = 0$ *and* $\omega(y, e, a, v)$ *does not depend on* e*; (iv)* V_{it} *is scalar and* $\omega(y, a, v) \equiv \bar{\omega}(y, a) + v$ *for some function* $\bar{\omega}$ *that is increasing in both* y *and* a*, and supermodular in* (y, a)*. Then Assumption MTS is satisfied with any value of* q*.*

The intuitive interpretation of Assumption MTS is that agents who are employed tend to be so because they have a permanently higher latent propensity to be employed. The conditions in Proposition 9 achieve this by imposing enough structure to make A_i this propensity. In the current model, this requires taking a stand on the relative importance of unobserved heterogeneity for wage offers and the cost of search effort. A simple way to do this is to simply shut down the search effort choice, as in (iii). This is not a terribly attractive assumption, since it removes the endogenous source of state dependence in the model, although an exogenous source of state dependence still exists via the wage offer function. For this reason, Assumption MTS will only be used in a secondary capacity in the next section.

Faberman et al. (2017) find survey evidence that the distribution of wage offers for employed workers dominates that of non-employed workers. The next assumptions capture this idea in differing strengths. To state them, let $e^\star(S_{i(t-1)} \| y)$ denote the optimal choice of effort in period $t-1$, given employment choice y, and let $W_{it}(y) \equiv \omega(y, e^\star(S_{i(t-1)} \| y), A_i, V_{it})$ denote the wage that would have been received in period t under these choices.

Assumption W(d): *The distribution of* $W_{it}(1)$ *first order stochastically dominates that of* $W_{it}(0)$*, conditional on* A_i*.*

Assumption W(e): $W_{it}(1)$ *is greater than* $W_{it}(0)$ *with probability* 1*.*

Assumption W(d) says that conditional on permanent heterogeneity, employed workers tend to receive higher wage offers than unemployed workers. Assumption W(e) strengthens this to assume that employed workers always receive better offers.

Assumptions W(d) and W(e) imply that the generated DPO model satisfies Assumptions MATR and the stronger Assumption MTR, respectively.

Proposition 10. *Suppose that every* $P \in \mathcal{P}^\dagger$ *is consistent with* $U_{it}(y)$ *being generated through* (7) *by the DC model described in this section. Suppose that Assumptions MC and W(b) are satisfied. If Assumption W(d) is satisfied, then Assumption MATR is satisfied. If Assumption W(e) is satisfied, then Assumption MTR is satisfied.*

There is ample reason to be skeptical of both Assumptions W(d) and W(e). For example, unemployed workers might receive higher offers than employed workers because they exert more effort searching for offers. Perhaps for this reason, the results in the

next section indicate that the model is misspecified under Assumption MTR. While there is no such evidence against Assumption MATR, this assumption turns out to have little identifying content, so will not be used in the main results.

5.4 Empirical Results

Table 2 reports estimated identified sets for a variety of parameters under several combinations of assumptions. The reported target parameters are time-averages of those discussed in Section 3.2, e.g.

$$\mathrm{SD}^+_{\mathrm{avg}}(P) \equiv \frac{1}{T}\sum_{t=1}^{T} \mathrm{SD}^+_t(P),$$

and similarly for the other parameters. For columns (1)–(8), the linear programs in Proposition 2 are feasible, so the estimated bounds are equal to the sharp identified set under the sample distribution of the data.[26] For columns (9)–(10), the programs were infeasible, so the estimated bounds are constructed using the estimation procedure described in Appendix G. For all columns, the reported 95% confidence intervals were constructed using the procedure of Chernozhukov, Newey, and Santos (2015), which is also described in Appendix G. The Monte Carlo results reported there suggest that these confidence intervals control size but are excessively wide.[27]

Column (1) contains the empirical-evidence-only bounds. These are wide for all target parameters, and in some cases completely uninformative. Consistent with the predictions of Proposition 1, the bounds for $\mathrm{SD}^+_{\mathrm{avg}}$ include 0 and the identified set for $\mathrm{SD}_{\mathrm{avg}}$ is $[0, 1]$. The upper bound on $\mathrm{SD}^+_{\mathrm{avg}}$ is large (.947), which reflects the strong positive serial correlation present in the observed data. To draw informative conclusions, more assumptions must be imposed.

Columns (2)–(6) impose Assumption ST for increasingly long sequence lengths, m. As expected, the bounds narrow as m increases, with the most informative bounds being obtained for $m = 4$, which is the strongest form of Assumption ST possible given the horizon of $T = 6$. For $m \geq 1$, the bounds on all target parameters exclude 0. In particular, the lower bound on overall state dependence, $\mathrm{SD}_{\mathrm{avg}}$, ranges between .018 and .061. This provides simple, nonparametric evidence against the hypothesis

[26] The linear programs were solved using AMPL (Fourer, Gay, and Kernighan, 2002) and CPLEX (IBM, 2010).

[27] Note that a tuning parameter, called τ_n in Appendix G, is required both for estimating identified sets when they are empty in sample, and for constructing confidence regions. I set this parameter to $\tau_n = .25$ throughout. This value was chosen because in the Monte Carlo simulation in Appendix G.6 it yields confidence regions with the correct coverage probability at the boundaries of the identified set.

that persistence in employment outcomes is caused solely by persistent unobserved heterogeneity. The small magnitude of these bounds is however still consistent with state dependence being of limited importance.

On the other hand, the results indicate that the role of state dependence is quite important for unemployed men. With $m = 2$ in column (4), the lower bound of .193 on $\text{SD}^+_{\text{avg}}(\cdot|0)$ means that at least 19.3% of unemployed men remain unemployed in the subsequent period due to state dependence, that is, due to the fact that they are unemployed. Using the interpretation discussed in Section 3.2, the lower bound of .335 on $\text{SD}^+_{\text{avg}}(\cdot|00)$ means that this causal effect accounts for at least 33.5% of the four-month persistence in unemployment. With $m = 4$ in column (6), these numbers rise to 23.8% and 41.4%, respectively.

In contrast, the bounds under Assumption ST alone are not very informative for employed men. This changes when Assumption MTS is added in columns (7)–(8), with conditioning length $q = 2$ or 3. These specifications generate informative upper bounds on state dependence among employed workers. In particular, the results in column (7) imply that no more than 37.1% of employed workers remain employed due to state dependence, and that state dependence accounts for no more than 38.2% of the four-month persistence in employment. Intuitively, these bounds arise because Assumption MTS reflects an assumption of some persistent heterogeneity, which caps the role of state dependence in explaining persistence in observed outcomes.

Column (9) adds Assumption MTR to Assumption ST with $m = 2$. Doing so causes the sample sharp identified set to become empty. This suggests misspecification, and a formal misspecification test provides evidence against the null of correct specification with a p-value of .012.[28] Such a finding is perhaps unsurprising given the strong assumptions required to motivate Assumption MTR in the job search model. Even if we were to put aside these concerns, adding Assumption MTR does little to narrow the bounds relative to column (4), so there would be little benefit to considering it anyway.[29]

Column (10) reports estimates that use the restrictions implied by Assumption TIV in Proposition 5. The sample identified set is empty, and a formal misspecification test rejects the null of correct specification with a p-value near 0. Like Assumption MTR,

[28] As suggested by a referee, the confidence regions may be too narrow under misspecification. Moreover, the estimated bounds will not be consistent if the model is actually misspecified, as suggested by these tests.

[29] Note that since the bounds in column (9) are estimated using the procedure in Appendix G.2, they need not be narrower than those in column (4), even though column (4) maintains fewer assumptions. For example, the lower bound on SD_{avg} is larger in column (4) than in column (9). If the population identified sets were non-empty for both specifications, then asymptotically the bounds in column (9) would become subsets of those in column (4).

Table 2: Nonparametric Estimates of State Dependence in Unemployment

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
					Assumptions					
ST(m)		0	1	2	3	4	4	4	2	
MTS(q)							2	3		
MTR									✓	
TIV										✓
					Misspecification					
$\Theta^\star = \emptyset$ in sample	No	No	No	No	No	No	No	No	Yes	Yes
p-value for $H_0 : \Theta^\star \neq \emptyset$									.012	.000
					Bounds and 95% Confidence Intervals					
SD^+_{avg}	.000 .000 .947 .950	.000 .000 .933 .939	.005 .016 .933 .939	.011 .025 .933 .939	.013 .030 .933 .939	.013 .034 .933 .939	.013 .034 .379 .468	.013 .034 .372 .451	.036 .037 .932 .935	.027 .028 .904 .904
$SD^+_{avg}(\cdot\|0)$	.000 .000 .581 .641	.000 .000 .581 .641	.030 .096 .581 .641	.065 .193 .576 .640	.072 .214 .574 .643	.088 .238 .569 .649	.088 .238 .554 .651	.088 .238 .544 .650	.217 .218 .514 .515	.164 .164 .581 .582
$SD^+_{avg}(\cdot\|00)$	.000 .000 1.00 1.00	.000 .000 1.00 1.00	.054 .166 1.00 1.00	.114 .335 .992 1.00	.128 .372 .990 1.00	.153 .414 .980 1.00	.153 .414 .956 1.00	.153 .414 .938 1.00	.379 .380 .891 .892	.286 .286 1.00 1.00
$SD^+_{avg}(\cdot\|1)$	.000 .000 .970 .975	.000 .000 .970 .975	.002 .006 .968 .974	.003 .010 .966 .973	.006 .014 .964 .972	.006 .016 .963 .971	.006 .016 .371 .463	.006 .016 .364 .446	.021 .021 .961 .964	.016 .016 .926 .926
$SD^+_{avg}(\cdot\|11)$	.000 .000 1.00 1.00	.000 .000 1.00 1.00	.002 .006 .998 1.00	.004 .010 .996 1.00	.006 .014 .994 1.00	.006 .017 .993 1.00	.006 .017 .382 .477	.006 .017 .376 .460	.021 .022 .991 .994	.016 .017 .955 .955
SD_{avg}	.000 .000 1.00 1.00	.013 .018 .976 .983	.023 .036 .976 .983	.035 .054 .976 .984	.035 .058 .976 .985	.034 .061 .975 .986	.034 .061 .422 .511	.034 .061 .412 .496	.036 .037 .932 .935	.029 .030 .940 .940

Notes: Estimated bounds and 95% confidence intervals are reported in large and small font respectively. Confidence intervals and misspecification p-values are obtained using the CNS method discussed in Appendices G.3–G.5 with 250 bootstrap draws and $\tau_n = .25$. In cases where the identified set is empty in sample, an estimate of the identified set is constructed using the method discussed in Appendix G.2. When the identified set is non-empty in sample, the p-value is of the CNS misspecification test is 1, and therefore not reported—see Appendix G.4.

Assumption TIV was also difficult to justify in the job search model, so this finding may not be surprising. The estimated identified sets in column (10) are similar to those obtained under Assumption ST with $m = 2$ in column (4), and suggest in particular that state dependence is important for explaining persistence in unemployment. However, given the resounding rejection of the specification, it seems prudent not to put much weight on these estimates.

Table 2 does not include estimates under Assumptions DSC, MATR, or MIV. Assumptions DSC and MATR turned out to have very little impact on the bounds in this application. Estimated identified sets for these assumptions are reported in Appendix H. Assumption MIV requires a compelling instrumental variable. This is not easy to find for the current application, but may be easier to find in other settings.[30]

5.5 Sensitivity to Stationarity

The results in the previous section show that Assumption ST is sufficient to reach the conclusion that state dependence is important for unemployed workers. Stationarity is a widely used assumption, and was easy to motivate in the job search model in Section 5.3. However, it may fail to hold if the causal mechanisms underlying the labor market changed over the sample of April 2011 to January 2013, or if there were seasonal changes within this horizon. It could also fail to the extent that the infinite horizon assumption ($\bar{T} = \infty$) used in Section 5.3 fails as a model of dynamic choice behavior. In this section, I examine the sensitivity to violations of Assumption ST.

To do so, I replace Assumption ST to the following, weaker condition.

Assumption ST(σ)***:*** *Let $\sigma \geq 0$ be a known scalar. Then for any $P \in \mathcal{P}^{\dagger}$, every $u \in \{0,1\}^{2(m+1)}$, and every $t, t' \geq m+1$,*

$$(1-\sigma)\Sigma^u_{t',m}(P) \leq \Sigma^u_{t,m}(P) \leq (1+\sigma)\Sigma^u_{t',m}(P),$$

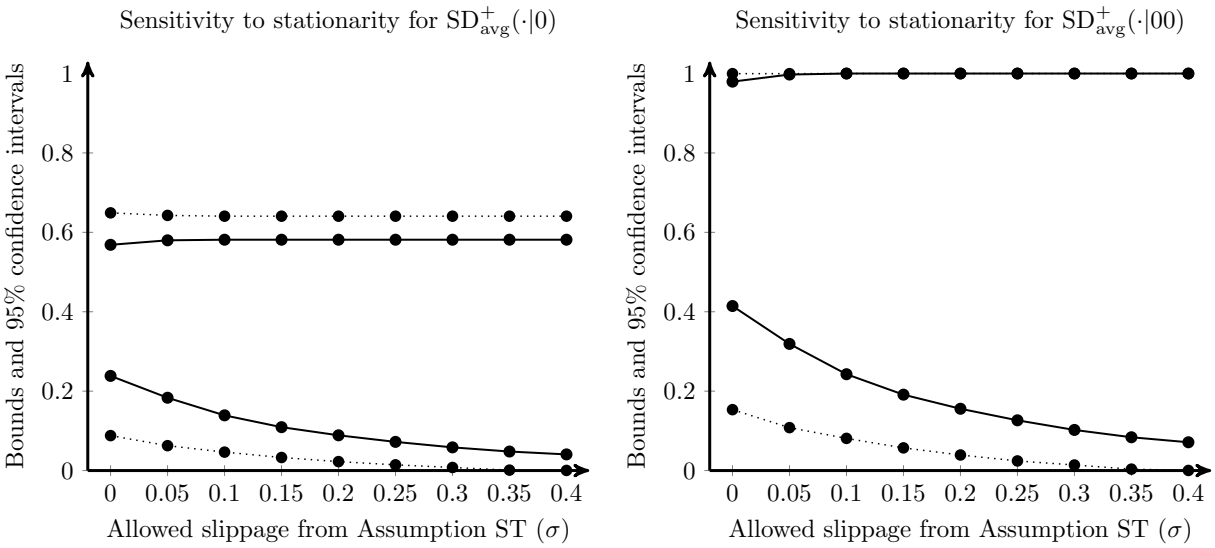
where $\Sigma^u_{t,m}$ is defined as in Assumption ST.

The parameter σ can be interpreted as the amount of "slippage" in $\Sigma_{t,m}(u;P)$ that is allowed in any two periods. For $\sigma = 0$, Assumption ST(σ) is the same as Assumption ST. For $\sigma > 0$, Assumption ST(σ) is strictly weaker than Assumption ST, since it allows the counterfactual probabilities to change by up to $100 \times \sigma\%$ in any two periods.

Figure 2 plots estimated identified sets and 95% confidence regions for $\text{SD}^{+}_{\text{avg}}(\cdot|0)$ and $\text{SD}^{+}_{\text{avg}}(\cdot|00)$ across different values of σ. The bounds widen as σ increases, reflect-

[30] One possibility, suggested by a referee, is to use variation in unemployment insurance benefits as in Farber, Rothstein, and Valletta (2015) or Farber and Valletta (2015).

Figure 2: Sensitivity of Table 2, Column (6) to Assumption ST

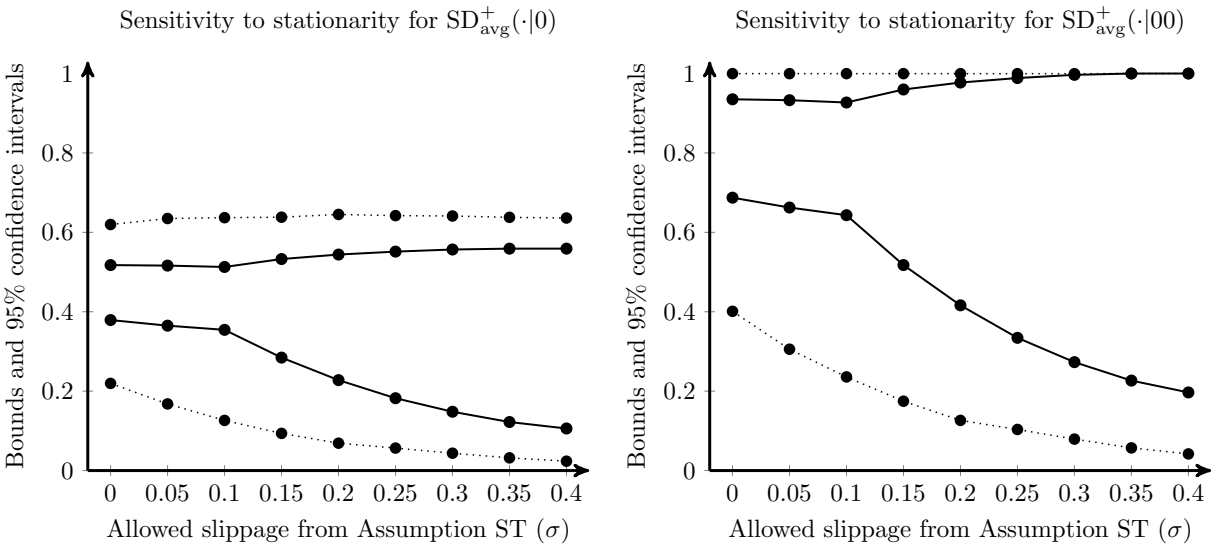


ing the fact that the restriction of Assumption ST(σ) becomes monotonically weaker. A value for σ of .15 means that the distribution of potential employment outcomes between two periods is allowed to change by up to 15%. This would reflect an enormous change in the structure of the labor market. Yet, even for values of σ larger than this, the lower bounds on $SD^{+}_{avg}(\cdot|0)$ and $SD^{+}_{avg}(\cdot|00)$ are substantial, indicating that state dependence has an important impact on the unemployed. Thus, the main conclusions regarding the unemployed appear quite robust.

As a further check, I estimate the same bounds for a sample of workers aged 40 or younger in the initial period. For these workers, the infinite horizon assumption is more reasonable, since they are farther from retirement. The estimated identified sets and confidence regions are plotted in Figure 3.[31] The lower bound estimates are substantially larger than those in Figure 2, and remain quite large even at values of σ that would represent massive changes in the underlying labor market. For younger workers, the conclusion that state dependence is important for the unemployed appears to be even stronger, even while this sample is also more likely to satisfy Assumption ST.

[31]Note that the analog identified sets are empty in the sample here, so the bounds are constructed using the estimation procedure discussed in Appendix G.2. This explains the slight decline in the upper bound from $\sigma = .05$ to $\sigma = .1$. The p-values for testing the null of correct specification are larger than .55 for all values of σ.

Figure 3: Estimates for Younger Workers under Assumption ST(σ) with $m = 4$



6 Conclusion

In this paper, I have developed a dynamic potential outcomes (DPO) model as a tool for empirically measuring state dependence in dynamic discrete outcomes. The DPO model has the important advantage of being fully nonparametric. Its primary disadvantage is that causal parameters tend to only be partially identified. Nevertheless, in applying the method to study state dependence in unemployment using data from the SIPP, I demonstrated that the estimated identified sets can still be tight enough to be useful. In particular, those estimates provide nonparametric evidence using observational data that state dependence is an important phenomenon in the employment dynamics of working age men in the U.S. The estimates rest on a stationarity assumption that the underlying economic environment remains stable, or least does not change drastically.

A Extension to Discrete Outcomes

The DPO model extends readily to the case where Y_{it} assumes values in $\{0, 1, \ldots, J\}$ with $J > 1$, so that assumes values in $\mathcal{Y} \equiv \{0, 1, \ldots, J\}^{T+1}$. Applications of such an extension to the dynamics of employment include Magnac (2000) and Prowse (2012), who examine state dependence under finer categorizations (part-time, full-time, etc.) of employment status. Irace (2018) applies the DPO model with $J > 1$ to the dynamics of hospital choice.

In this more general case, there are $J + 1$ potential outcomes $\{U_{it}(y)\}_{y=0}^{J}$ for each $t \geq 1$. The observed outcome in period t is determined as

$$Y_{it} = \sum_{y=0}^{J} \mathbb{1}[Y_{i(t-1)} = y] U_{it}(y).$$

The primitive P is a probability mass function for $(Y_{i0}, \{U_{it}(0), \ldots, U_{it}(J)\}_{t=1}^{T})$ and the characterization of the identified set remains conceptually unchanged. Some parameters and identifying assumptions that are appropriate for the $J = 1$ case are also appropriate for the $J > 1$ case, but others would require modification. A separate analysis seems beyond the scope of this paper.

B Extension to Higher Order State Dependence

The discussion in the main text presumes that the analyst is interested in first order state dependence, i.e. the causal effect of the immediately preceding period on the current period. This is consistent with much of the empirical and theoretical literature on discrete state dependence. However, the DPO model can also be modified to consider the causal effects of longer histories of past outcomes. In this section, I outline how one would extend the model to enable the estimation of this type of higher order state dependence.

When Y_t is binary, the generalization to state dependence of length $K \geq 1$ is accomplished by introducing 2^K potential outcomes $\{U_{it}(y)\}_{y \in \{0,1\}^K}$ for each period $t \geq K$. The recursive relationship (1) is replaced by

$$Y_{it} = \sum_{y \in \{0,1\}^K} U_{it}(y) \mathbb{1}[(Y_{it}, Y_{i(t-1)}, \ldots, Y_{i(t-K+1)}) = y] \quad \text{for } t \geq K, \tag{25}$$

with the joint determination of periods $t = 0$ up to $t = K - 1$ not being modeled explicitly. For example, with $K = 2$, (25) would become

$$\begin{aligned} Y_{it} = {} & \mathbb{1}[Y_{i(t-1)} = 0, Y_{i(t-2)} = 0] U_{it}(0,0) + \mathbb{1}[Y_{i(t-1)} = 0, Y_{i(t-2)} = 1] U_{it}(0,1) \\ & + \mathbb{1}[Y_{i(t-1)} = 1, Y_{i(t-2)} = 0] U_{it}(1,0) + \mathbb{1}[Y_{i(t-1)} = 1, Y_{i(t-2)} = 1] U_{it}(1,1), \end{aligned}$$

so that for each t there are four potential outcomes corresponding to the four potential two-period histories immediately prior to period t. The primitive P is a probability

mass function for the random vector

$$\left(Y_{i0}, Y_{i1}, \ldots, Y_{i(K-1)}, \left\{U_{it}(y) : y \in \{0,1\}^K\right\}_{t=K}^{T}\right).$$

The identified set $\mathcal{P}^\star$ can be characterized through essentially the same argument as for the first-order case.

C A Brief Survey of Semiparametric Identification in DBR Models

The most common implementation of (8) constructs a finitely parameterized likelihood function by imposing a parametric distributional assumption (typically normality) for both $(V_{i1}, \ldots, V_{iT})$ and A_i, and by further assuming that these latent variables are independent of X_i, as well as independent of each other. Maximum likelihood estimates of the parameters in (8) can be used with the maintained parametric distributional assumptions to form estimates of causal parameters like those discussed in Section 3. However, consistency of these estimates depends critically on the validity of the parametric assumptions.

Honoré (2002) raised a number of additional criticisms of this reduced form approach. Many of his points apply equally to parametric implementations of structural DC models like (2). One frequently discussed criticism is the treatment of A_i as a normally distributed random effect that is independent of other explanatory state variables. Since (8) is nonlinear, treating A_i as a fixed effect to be estimated leads to the well-known incidental parameters problem when T is small.[32] Nonlinear differencing arguments can be applied in certain cases (Chamberlain 1984; 1985; 2010, Honoré and Kyriazidou 2000, and Bartolucci and Nigro 2010), but these depend on very specific functional forms and may therefore amplify concerns about misspecification.[33] Honoré and Lewbel (2002) showed that if there exists an exogenous special regressor with a large amount of variation, then A_i can be treated as a fixed effect while also relaxing distributional assumptions in (8). However, their results only identify the parameter coefficients and not causal parameters such as the ATE. Similarly, Pakes and Porter (2016) allow A_i to be a fixed effect and also remove parametric distributional assumptions on V_{it}, but their partial identification results concern the index parameters, not the causal parameters that constitute the focus of this paper.

Some researchers have developed point and partial identification results for nonparametric models of dynamic binary outcomes that depart from the threshold crossing form of (8) entirely. Instead, these papers maintain assumptions that imply that Y_{it} follows a homogenous, first-order Markov process, conditional on permanent latent variable, like A_i, and possibly also on the initial period, Y_{i0}. Under this type of assumption, Kasahara and Shimotsu (2009) and Browning and Carro (2010, 2014), establish point identification by imposing the additional assumption that A_i has sufficiently small finite support relative to T, while Hu and Shum (2012) allow A_i to be continuously distributed, but impose a high-level completeness condition. These

[32] Fernández-Val (2009) argues that the bias on the ATE may be relatively small, even for small T, while Carro (2007) argues that the bias can be mitigated by using a modified maximum likelihood estimator.

[33] See also Bonhomme (2012) for related results and a unifying analysis.

types of conditions on the distribution of unobserved heterogeneity may be difficult to motivate or interpret in applications. In addition, the first-order conditional Markov property assumed by all of these papers may be unattractive in some settings (see Bhuller, Brinch, and Konigs (2016) and Section 6 of Browning and Carro (2014)). The DPO model does not maintain this type of first-order conditional Markov property, although it can be imposed if desired; see Section 4.4.

D Proofs

Proof of Proposition 1. Observe that if $P \in \mathcal{P}^\star$ then

$$\begin{aligned}
\mathrm{SD}_t^+(P) &= \mathbb{P}_P[Y_{i(t-1)} = 0, U_{it}(0) = 0, U_{it}(1) = 1] \\
&\quad + \mathbb{P}_P[Y_{i(t-1)} = 1, U_{it}(0) = 0, U_{it}(1) = 1] \\
&= \mathbb{P}_P[Y_{i(t-1)} = 0, Y_{it} = 0, U_{it}(1) = 1] + \mathbb{P}_P[Y_{i(t-1)} = 1, U_{it}(0) = 0, Y_{it} = 1] \\
&= \mathbb{P}[Y_{i(t-1)} = 0, Y_{it} = 0] + \mathbb{P}[Y_{i(t-1)} = 1, Y_{it} = 1] \\
&\quad - \mathbb{P}_P[Y_{i(t-1)} = 0, Y_{it} = 0, U_{it}(1) = 0] - \mathbb{P}_P[Y_{i(t-1)} = 1, U_{it}(0) = 1, Y_{it} = 1],
\end{aligned}$$

where the second equality follows because under (1), $[Y_{i(t-1)} = 0, U_{it}(0) = 0]$ if and only if $[Y_{i(t-1)} = 0, Y_{it} = 0]$, and $[Y_{i(t-1)} = 1, U_{it}(1) = 1]$ if and only if $[Y_{i(t-1)} = 1, Y_{it} = 1]$. The only restrictions implied on the second two terms are

$$\begin{aligned}
0 \geq &- \mathbb{P}_P[Y_{i(t-1)} = 0, Y_{it} = 0, U_{it}(1) = 0] \geq - \mathbb{P}_P[Y_{i(t-1)} = 0, Y_{it} = 0] \\
0 \geq &- \mathbb{P}_P[Y_{i(t-1)} = 1, U_{it}(0) = 1, Y_{it} = 1] \geq - \mathbb{P}_P[Y_{i(t-1)} = 1, Y_{it} = 1], \qquad (26)
\end{aligned}$$

and there are no cross-equation restrictions between these terms. Hence there exists a $P \in \mathcal{P}^\star$ obtaining both of the upper bounds in (26), and one obtaining both of the lower bounds. The upper and lower bounds in (14) now follow from those in (26). The bounds in (15) follow from an analogous argument using the decomposition

$$\begin{aligned}
\mathrm{SD}_t^-(P) &= \mathbb{P}[Y_{i(t-1)} = 0, Y_{it} = 1] + \mathbb{P}[Y_{i(t-1)} = 1, Y_{it} = 0] \\
&\quad - \mathbb{P}_P[Y_{i(t-1)} = 0, Y_{it} = 1, U_{it}(1) = 1] - \mathbb{P}_P[Y_{i(t-1)} = 1, U_{it}(0) = 0, Y_{it} = 0],
\end{aligned}$$

and the following analogous bounds on the second two terms

$$\begin{aligned}
0 \geq &- \mathbb{P}_P[Y_{i(t-1)} = 0, Y_{it} = 1, U_{it}(1) = 1] \geq - \mathbb{P}_P[Y_{i(t-1)} = 0, Y_{it} = 1] \\
0 \geq &- \mathbb{P}_P[Y_{i(t-1)} = 1, U_{it}(0) = 0, Y_{it} = 0] \geq - \mathbb{P}_P[Y_{i(t-1)} = 1, Y_{it} = 0]. \qquad (27)
\end{aligned}$$

Observe that there are no restrictions preventing the terms in (26) and (27) from simultaneously achieving either their lower or upper bounds. Considering these two polar cases shows that the sharp identified set for $\mathrm{SD}_t \equiv \mathrm{SD}_t^+ + \mathrm{SD}_t^-$ is equal to $[0, 1]$. *Q.E.D.*

Proof of Proposition 2. If $\mathcal{P}^\dagger$ is closed and convex, then $\mathcal{P}^\star$ is also closed and convex, since $\mathcal{P}^\star$ is the set of $P \in \mathcal{P}^\dagger$ that satisfy the linear equalities (11) for all y and x. The image of the continuous, real-valued function θ over this closed, convex, and

non-empty set is a closed, non-empty interval (e.g. Rudin, 1976, Theorem 4.22) with smallest value $\theta^\star_{\text{lb}}$ and largest value $\theta^\star_{\text{ub}}$, i.e. $\Theta^\star = [\theta^\star_{\text{lb}}, \theta^\star_{\text{ub}}]$. *Q.E.D.*

Proof of Proposition 3. Let $P \in \mathcal{P}$ denote a distribution that is consistent with the stated conditions and fix $y \in \{0, 1\}$. Since $\varphi(\bar{s}, \cdot)$ is weakly increasing for each $\bar{s}$, it has a generalized inverse, $\varphi^{-1}(\bar{s}, \cdot)$. By Theorem 3.1 of Fang, Hu, and Joe (1994), $(\tilde{S}_{it}, \tilde{S}_{i(t+t')})$ is decreasing in the concordance ordering as a function of $|t'|$, conditional on $\bar{S}_i$. That is, $\mathbb{P}[\tilde{S}_{it} \geq \tilde{s}_1, \tilde{S}_{i(t+t')} \geq \tilde{s}_2 | \bar{S}_i = \bar{s}]$ is decreasing in $|t'|$ for any $\tilde{s}_1, \tilde{s}_2$ and $\bar{s}$. Thus, for any integers t', t'' with $|t'| < |t''|$

$$\begin{aligned}
&\mathbb{P}_P[U_{it}(y) = 1, U_{i(t+t')}(y) = 1] \\
&\quad = \mathbb{P}\left[\Delta\mathring{\nu}(S_{it}(y)) \geq 0,\ \Delta\mathring{\nu}(S_{i(t+t')}(y)) \geq 0\right] \\
&\quad = \mathbb{E}\left[\mathbb{P}\left[\varphi(\bar{S}_i, \tilde{S}_{it}) \geq 0,\ \varphi(\bar{S}_i, \tilde{S}_{i(t+t')}) \geq 0 | \bar{S}_i\right]\right] \\
&\quad = \mathbb{E}\left[\mathbb{P}\left[\tilde{S}_{it} \geq \varphi^{-1}(\bar{S}_i, 0),\ \tilde{S}_{i(t+t')} \geq \varphi^{-1}(\bar{S}_i, 0) | \bar{S}_i\right]\right] \\
&\quad \geq \mathbb{E}\left[\mathbb{P}\left[\tilde{S}_{it} \geq \varphi^{-1}(\bar{S}_i, 0),\ \tilde{S}_{i(t+t'')} \geq \varphi^{-1}(\bar{S}_i, 0) | \bar{S}_i\right]\right] \\
&\quad = \mathbb{P}_P[U_{it}(y) = 1, U_{i(t+t'')}(y) = 1],
\end{aligned} \tag{28}$$

where the third equality follows because $\varphi(\bar{s}, \cdot)$ is right-continuous (e.g. Embrechts and Hofert, 2013, Proposition 1(5)), and the final equality reverses the steps of the first three. As discussed in Appendix E, (28) implies Assumption DSC if Assumption ST is satisfied. *Q.E.D.*

Proof of Proposition 4. Let $P \in \mathcal{P}$ denote a distribution that is consistent with the stated conditions. If A and B are two events and B occurs with probability strictly between 0 and 1, then $\mathbb{P}[A, B] \geq \mathbb{P}[A]\,\mathbb{P}[B]$ implies $\mathbb{P}[A|B] \geq \mathbb{P}[A|B^c]$.[34] From this, (19) follows from conditions (i) and (ii) whenever $\mathbb{P}[Y_{i(t-1)} = 1, Y_{i(t-2)} = \tilde{y}] \in (0, 1)$:

$$\begin{aligned}
&\mathbb{P}_P[U_{it}(y) = 1 | Y_{i(t-1)} = 1, Y_{i(t-2)} = \tilde{y}] \\
&\quad = \mathbb{P}_P[\Delta\mathring{\nu}(S_{it}(y)) \geq 0 | \Delta\mathring{\nu}(S_{i(t-1)}(\tilde{y})) \geq 0, Y_{i(t-2)} = \tilde{y}] \\
&\quad \geq \mathbb{P}_P[\Delta\mathring{\nu}(S_{it}(y)) \geq 0 | \Delta\mathring{\nu}(S_{i(t-1)}(\tilde{y})) < 0, Y_{i(t-2)} = \tilde{y}] \\
&\quad = \mathbb{P}_P[U_{it}(y) = 1 | Y_{i(t-1)} = 0, Y_{i(t-2)} = \tilde{y}].
\end{aligned}$$

Q.E.D.

Proof of Proposition 5. Note that Assumption TIV implies that

$$\mathbb{P}[U_{it} = u | Y_{i(0:t-1)}, A_i] = \mathbb{P}[U_{i1} = u | Y_{i0}, A_i] = \mathbb{P}[U_{it'} = u | Y_{i(0:t'-1)}, A_i]$$

[34] This follows because $\mathbb{P}\left[A \mid B\right] \equiv \mathbb{P}[A, B]\,\mathbb{P}[B]^{-1} \geq \mathbb{P}[A]$, while on the other hand $\mathbb{P}\left[A \mid B^c\right] = (\mathbb{P}[A] - \mathbb{P}[A, B])(1 - \mathbb{P}[B])^{-1} \leq \mathbb{P}[A]$.

almost surely, for any t, t' and $u \in \{0,1\}^2$. As a consequence, if $t' > t \geq 1$ then

$$\begin{aligned}
&\mathbb{P}_P\left[U_{it'} = u, Y_{i(0:t-1)} = y\right] \\
&\quad = \mathbb{E}\left[\mathbb{P}\left[U_{it'} = u, Y_{i(0:t-1)} = y | Y_{i(0:t'-1)}, A_i\right]\right] \\
&\quad = \mathbb{E}\left[\mathbb{1}\left[Y_{i(0:t-1)} = y\right]\mathbb{P}\left[U_{it'} = u | Y_{i(0:t'-1)}, A_i\right]\right] \\
&\quad = \mathbb{E}\left[\mathbb{1}\left[Y_{i(0:t-1)} = y\right]\mathbb{P}\left[U_{it} = u | Y_{i(0:t-1)}, A_i\right]\right] \\
&\quad = \mathbb{E}\left[\mathbb{P}\left[U_{it} = u, Y_{i(0:t-1)} = y | Y_{i(0:t-1)}, A_i\right]\right] \\
&\quad = \mathbb{P}_P\left[U_{it} = u, Y_{i(0:t-1)} = y\right],
\end{aligned}$$

for any $u \in \{0,1\}^2$ and $y \in \{0,1\}^t$, as claimed. Summing both sides of this equality over all realizations y of $Y_{i(0:t-1)}$ shows that Assumption TIV implies Assumption ST with $m = 0$.

Q.E.D.

Proof of Proposition 6. Let $P \in \mathcal{P}$ denote a distribution that is consistent with the stated conditions. Then for any $u = (u_0, u_1) \in \{0,1\}^2$,

$$\begin{aligned}
&\mathbb{P}_P\left[U_{it} = u \middle| Y_{i(0:t-1)}, \bar{S}_i\right] \\
&\quad = \mathbb{E}_P\left[\mathbb{P}_P\left[U_{it} = u \middle| Y_{i(0:t-1)}, \bar{S}_i, \tilde{S}_{i(1:t-1)}\right] \middle| Y_{i(0:t-1)}, \bar{S}_i\right] \\
&\quad = \mathbb{E}_P\left[\mathbb{P}_P\left[U_{it} = u \middle| Y_{i0}, \bar{S}_i, \tilde{S}_{i(1:t-1)}\right] \middle| Y_{i(0:t-1)}, \bar{S}_i\right] \\
&\quad = \mathbb{E}_P\left[\mathbb{P}_P\left[U_{i1} = u \middle| Y_{i0}, \bar{S}_i\right] \middle| Y_{i(0:t-1)}, \bar{S}_i\right] = \mathbb{P}_P\left[U_{i1} = u \middle| Y_{i0}, \bar{S}_i\right],
\end{aligned}$$

where the second equality follows because $Y_{i(1:t-1)}$ are fully determined by Y_{i0}, $\tilde{S}_{i(1:t-1)}$, and $\bar{S}_i$ under (7) and (1). The third equality used condition (ii), since when U_{it} is determined by (7), it is a function of $(S_{it}(0), S_{it}(1))$, which have here been split into $(\bar{S}_i, \tilde{S}_{it})$. The time-invariant variable $\bar{S}_i$ serves the role of A_i in the statement of Assumption TIV. *Q.E.D.*

Proof of Proposition 7. Let $P \in \mathcal{P}$ denote a distribution that is consistent with the stated conditions and fix a $y \in \{0,1\}$. Consider any $(x^0, x^1), (x^0, \tilde{x}^1)$ in the support of (X_{it}^0, X_{it}^1), with $\tilde{x}^1 \geq x^1$. Then

$$\begin{aligned}
&\mathbb{P}_P\left[U_{it}(y) = 1 | X_{it}^0 = x^0, X_{it} = x^1\right] \\
&\quad = \mathbb{P}_P\left[\varphi(x^0, x^1, V_{it}) \geq 0 | X_{it}^0 = x^0, X_{it}^1 = x^1\right] \\
&\quad = \mathbb{P}_P\left[\varphi(x^0, x^1, V_{it}) \geq 0 | X_{it}^0 = x^0, X_{it}^1 = \tilde{x}^1\right] \\
&\quad \leq \mathbb{P}_P\left[\varphi(x^0, \tilde{x}^1, V_{it}) \geq 0 | X_{it}^0 = x^0, X_{it}^1 = \tilde{x}^1\right] \\
&\quad = \mathbb{P}_P\left[U_{it}(y) = 1 | X_{it}^0 = x_0, X_{it}^1 = \tilde{x}_1\right],
\end{aligned}$$

where the second and third equalities used conditions (iii) and (iv). *Q.E.D.*

Proof of Proposition 8. Under Assumption MC, the worker's Bellman equation is given by (24). Let $\mathring{\nu}$ denote the worker's per-period objective function, as in (2).

Profiling the effort decision for a fixed employment decision y' gives

$$\begin{aligned} e^\star(S_{it}\|y') &\equiv \arg\max_{e'\in\mathcal{E}} \ \mathring{\nu}\left(y', e', Y_{i(t-1)}, E_{i(t-1)}, V_{it}, A_i\right) \\ &= \arg\max_{e'\in\mathcal{E}} \ -\kappa(y', e', A_i) + \delta\mathbb{E}\left[\nu\left(y', e', V_{i(t+1)}, A_i\right) \mid V_{it}, A_i\right], \end{aligned} \tag{29}$$

so that the effort decision in period t is a function of the (fixed) current period employment decision, y', the current period wage shock, V_{it}, and time-invariant heterogeneity, A_i. As a consequence, the counterfactual state in period t had the worker chosen y in period $t-1$, is

$$S_{it}(y) \equiv \left(y, e^\star\left(V_{i(t-1)}, A_i\|y\right), V_{it}, A_i\right),$$

which depends on the hypothesized previous period employment decision, y, the previous and current period wage shocks, $(V_{i(t-1)}, V_{it})$, and time-invariant heterogeneity, A_i.[35] The worker's present-discounted net utility from choosing employment if the previous period's employment choice was y can therefore be written as

$$\Delta\mathring{\nu}(S_{it}(y)) = \omega\left(y, e^\star(V_{i(t-1)}, A_i\|y), A_i, V_{it}\right) - \Delta\kappa\left(V_{it}, A_i\right) + \Delta\gamma\left(V_{it}, A_i\right), \tag{30}$$

where $\Delta\kappa$ and $\Delta\gamma$ are shorthand for

$$\begin{aligned} &\Delta\kappa\left(V_{it}, A_i\right) = \left[\kappa\left(1, e^\star(V_{it}, A_i\|1), A_i\right) - \kappa\left(0, e^\star(V_{it}, A_i\|0), A_i\right)\right], \\ \text{and} \quad &\Delta\gamma\left(V_{it}, A_i\right) = \delta\,\mathbb{E}\Big[\nu\left(1, e^\star(V_{it}, A_i\|1), V_{i(t+1)}, A_i\right) \\ &\qquad\qquad - \nu\left(0, e^\star(V_{it}, A_i\|0), V_{i(t+1)}, A_i\right)\Big| V_{it}, A_i\Big]. \end{aligned}$$

Statements (i)–(iv) can now be proven for the generated potential outcomes (7) by using (30) as follows.

(i) Notice that $(U_{it}(0), U_{it}(1))$ is fully determined by $(\Delta\mathring{\nu}(S_{it}(0)), \Delta\mathring{\nu}(S_{it}(1)))$, and that from (30), the latter is only stochastic due to $(V_{i(t-1)}, V_{it})$ and A_i). Thus, if Assumption W(a) is satisfied with $m' = 1$, then since A_i is time-invariant, Assumption ST is also satisfied with $m = m' - 1 = 0$. More generally, (7) and (30) imply that $U_{i(t-m:t)}(0)$ and $U_{i(t-m:t)}(1)$ are stochastic only due to $V_{i(t-m-1:t)}$ and A_i, so that Assumption W(a) with any $m' \geq 0$ implies Assumption ST with $m = m' - 1$.

(ii) Under Assumption W(b),

$$\mathbb{E}\left[\nu\left(y', e', V_{i(t+1)}, A_i\right) \mid V_{it}, A_i\right] = \mathbb{E}\left[\nu\left(y', e', V_{i(t+1)}, A_i\right) \mid A_i\right].$$

From (29), it follows that $e^\star(V_{it}, A_i\|y) = e^\star(A_i\|y)$ is only stochastic due to A_i. This implies that $S_{it}(y) = (y, e^\star(A_i\|y), V_{it}, A_i)$ is only stochastic due to V_{it} and A_i, and that

[35] As in Section 2.2, I am assuming here that the solution to this effort decision is unique.

(30) can be written as

$$\Delta\mathring{\nu}(S_{it}(y)) = \omega\left(y, e^{\star}\left(A_i \| y\right), A_i, V_{it}\right) - \Delta\kappa\left(A_i\right) + \Delta\gamma\left(A_i\right). \tag{31}$$

To see that the conditions of Proposition 3 are satisfied, fix a y, take $\bar{S}_i \equiv A_i$ and take $\tilde{S}_{it} \equiv \omega(y, e^{\star}(A_i \| y), A_i, V_{it})$. Then (31) can be written as

$$\Delta\mathring{\nu}(S_{it}(y)) \equiv \tilde{S}_{it} - \Delta\kappa\left(\bar{S}_i\right) + \Delta\gamma\left(\bar{S}_i\right) \equiv \varphi(\bar{S}_i, \tilde{S}_{it}), \tag{32}$$

which is an increasing and continuous function of $\tilde{S}_{it}$. Since V_{it} and $V_{i(t-1)}$ are independent, conditional on A_i, so too are $\tilde{S}_{it}$ and $\tilde{S}_{i(t-1)}$, conditional on $\bar{S}_i$. Therefore, the stochastic increasing condition of Proposition 3 is also satisfied.

(iii) As shown in (ii), Assumptions MC and W(b) imply that $(U_{it}(0), U_{it}(1))$ is stochastic only due to V_{it} and A_i. Thus, the same argument as in (i) holds with $m' = m$.

(iv) As shown in (ii), Assumptions MC and W(b) imply that the stochastic components of $(S_{it}(0), S_{it}(1))$ are $(V_{it}, A_i) \equiv (V_{it}, \bar{A}_i, Y_{i0})$. Assumption MC further implies that $V_{it}|\{V_{it'}\}_{t'<t}, \bar{A}_i, Y_{i0}$ has the same distribution as $V_{it}|V_{i(t-1)}, \bar{A}_i, Y_{i0}$. By Assumption W(b), the latter has the same distribution as $V_{it}|\bar{A}_i, Y_{i0}$, which by Assumption W(a) has the same distribution as $V_{i1}|\bar{A}_i, Y_{i0}$. Thus, the conditions of Proposition 6 are satisfied with $\bar{S}_i \equiv \bar{A}_i$ and $\tilde{S}_{it} \equiv V_{it}$.

Q.E.D.

Proof of Proposition 9. Under Assumptions MC and W(b), the worker's net utility from employment can be written as (31); see Proposition 8(ii). For shorthand, define

$$\Omega(y, A_i) \equiv \Delta\kappa\left(A_i\right) - \Delta\gamma\left(A_i\right) - \bar{\omega}\left(y, e^{\star}(A_i \| y), A_i\right).$$

Then since $\omega(y, e, a, v) = \bar{\omega}(y, e, a) + v$,

$$\begin{aligned}
&\mathbb{P}\left[\Delta\mathring{\nu}(S_{it}(y)) \geq 0,\ \Delta\mathring{\nu}(S_{i(t-1)}(\tilde{y})) \geq 0 \,\middle|\, Y_{i(t-2)} = \tilde{y}\right] \\
&\quad = \mathbb{E}\left[\mathbb{P}\left[V_{it} \geq \Omega(y, A_i),\ V_{i(t-1)} \geq \Omega(\tilde{y}, A_i) \,\middle|\, Y_{i(t-2)} = \tilde{y}, A_i, \{V_{is}\}_{s\leq t-2}\right] \,\middle|\, Y_{i(t-2)} = \tilde{y}\right] \\
&\quad = \mathbb{E}\left[\mathbb{P}\left[V_{it} \geq \Omega(y, A_i),\ V_{i(t-1)} \geq \Omega(\tilde{y}, A_i) \,\middle|\, A_i, \{V_{is}\}_{s\leq t-2}\right] \,\middle|\, Y_{i(t-2)} = \tilde{y}\right] \\
&\quad = \mathbb{E}\left[\begin{matrix} \mathbb{P}\left[V_{it} \geq \Omega(y, A_i) \mid A_i\right] \\ \times \quad \mathbb{P}\left[V_{i(t-1)} \geq \Omega(\tilde{y}, A_i) \mid A_i\right] \end{matrix} \,\middle|\, Y_{i(t-2)} = \tilde{y}\right],
\end{aligned} \tag{33}$$

where the second equality follows because $Y_{i(t-2)}$ is fully determined by A_i and $\{V_{is}\}_{s\leq t-2}$ when A_i includes the initial conditions, $(Y_{i\underline{T}}, E_{i\underline{T}})$, and the third equality uses Assumption W(b). Since A_i and V_{it} are assumed to be independent,

$$\mathbb{P}\left[V_{it} \geq \Omega(y, A_i) \mid A_i = a\right] = \mathbb{P}\left[V_{it} \geq \Omega(y, a)\right] \equiv F_t(\Omega(y, a)),$$

so that (33) can be written as

$$\begin{aligned}&\mathbb{P}\left[\Delta\mathring{\nu}(S_{it}(y)) \geq 0,\ \Delta\mathring{\nu}(S_{i(t-1)}(\tilde{y})) \geq 0 \,\middle|\, Y_{i(t-2)} = \tilde{y}\right] \\ &\quad = \mathbb{E}\left[F_t\left(\Omega(y, A_i)\right) F_{t-1}\left(\Omega(\tilde{y}, A_i)\right) \,\middle|\, Y_{i(t-2)} = \tilde{y}\right].\end{aligned}$$

Below, it will be shown that $F_t(\Omega(y, a)$ and $F_{t-1}(\Omega(\tilde{y}, a))$ are both increasing functions of a. This implies that the covariance between $F_t(\Omega(y, A_i))$ and $F_{t-1}(\Omega(\tilde{y}, A_i))$ is positive conditional on $Y_{i(t-2)} = \tilde{y}$ (or on any other event), see e.g. Lehmann (1966). Thus,

$$\begin{aligned}&\mathbb{P}\left[\Delta\mathring{\nu}(S_{it}(y)) \geq 0,\ \Delta\mathring{\nu}(S_{i(t-1)}(\tilde{y})) \geq 0 \,\middle|\, Y_{i(t-2)} = \tilde{y}\right] \\ &\quad \geq \mathbb{E}\left[F_t\left(\Omega(y, A_i)\right) \,\middle|\, Y_{i(t-2)} = \tilde{y}\right] \mathbb{E}\left[F_{t-1}\left(\Omega(\tilde{y}, A_i)\right) \,\middle|\, Y_{i(t-2)} = \tilde{y}\right]. \qquad (34)\end{aligned}$$

By reversing the previous arguments, one has

$$\begin{aligned}\mathbb{E}\left[F_t\left(\Omega(y, A_i)\right) \,\middle|\, Y_{i(t-2)} = \tilde{y}\right] &= \mathbb{P}\left[\Delta\mathring{\nu}(S_{it}(y)) \geq 0 \,\middle|\, Y_{i(t-2)} = \tilde{y}\right], \\ \text{and} \quad \mathbb{E}\left[F_{t-1}\left(\Omega(\tilde{y}, A_i)\right) \,\middle|\, Y_{i(t-2)} = \tilde{y}\right] &= \mathbb{P}\left[\Delta\mathring{\nu}(S_{i(t-1)}(\tilde{y})) \geq 0 \,\middle|\, Y_{i(t-2)} = \tilde{y}\right],\end{aligned}$$

which upon substitution into (34) implies that $\Delta\mathring{\nu}(S_{it}(y))$ and $\Delta\mathring{\nu}(S_{i(t-1)}(\tilde{y}))$ are positively quadrant dependent, locally at $(0, 0)$ and conditional on $Y_{i(t-2)} = \tilde{y}$. The result then follows from Proposition 4.

It remains to be shown that $F_t(\Omega(y, a))$ and $F_{t-1}(\Omega(\tilde{y}, a))$ are both increasing functions of a. By definition, both $F_t(\cdot)$ and $F_{t-1}(\cdot)$ are decreasing functions, so it suffices to show that $\Omega(y, a)$ is also a decreasing function of a. For this, we remove the search effort decision from the model (condition (iii)), under which

$$\Omega(y, a) = -\Delta\gamma(a) - \bar{\omega}\left(y, a\right).$$

The second term is decreasing in a by condition (iv). As for the first term, under the given assumptions we have

$$\Delta\gamma(a) = \delta\, \mathbb{E}\left[\nu\left(1, V_{i(t+1)}, a\right) - \nu\left(0, V_{i(t+1)}, a\right)\right],$$

so to show that it is increasing in a (and therefore that $-\Delta\gamma(y)$ is decreasing in a), it suffices to show that $\nu(y, v, a)$ has the increasing differences (or supermodularity in this simple setting) property in (y, a), i.e. that $\nu(1, v, a) - \nu(0, v, a)$ is increasing as a function of a for all v. Under the maintained assumptions, the Bellman equation is

$$\nu(y, v, a) = \max_{y' \in \{0,1\}} \left\{y'\left(\bar{\omega}(y, a) + v\right) + \delta\, \mathbb{E}\left[\nu\left(y', V_{i(t+1)}, a\right)\right]\right\},$$

so that $\nu(y, v, a)$ will have increasing differences in (y, a) for all v if $\tilde{\omega}(y', y, a, v) \equiv y'(\bar{\omega}(y, a) + v)$ has increasing differences in (y', y) for all (a, v), in (y', a) for all (y, v), and in (y, a) for all (y', v); see Proposition 2 of Hopenhayn and Prescott (1992). To see that these conditions are satisfied here, observe that by condition (iv),

$$\tilde{\omega}(1, y, a, v) - \tilde{\omega}(0, y, a, v) = \bar{\omega}(y, a) + v$$

is increasing in both y and a, and

$$\tilde{\omega}(y', 1, a, v) - \tilde{\omega}(y', 0, a, v) = y' \left(\bar{\omega}(1, a) - \bar{\omega}(0, a)\right)$$

is increasing in a. *Q.E.D.*

Proof of Proposition 10. Under Assumptions MC and W(b), $\Delta\mathring{\nu}(S_{it}(y))$ is given by (31). Assumption W(d) then implies that

$$\begin{aligned}
&\mathbb{P}\left[\Delta\mathring{\nu}(S_{it}(0)) \geq 0\right] \\
&\quad = \mathbb{P}\left[W_{it}(0) \geq \Delta\kappa(A_i) - \Delta\gamma(A_i)\right] \\
&\quad = \mathbb{E}\left[\mathbb{P}\left[W_{it}(0) \geq \Delta\kappa(A_i) - \Delta\gamma(A_i) \mid A_i\right]\right] \\
&\quad \leq \mathbb{E}\left[\mathbb{P}\left[W_{it}(1) \geq \Delta\kappa(A_i) - \Delta\gamma(A_i) \mid A_i\right]\right] = \mathbb{P}\left[\Delta\mathring{\nu}(S_{it}(1)) \geq 0\right],
\end{aligned}$$

which implies Assumption MATR when $U_{it}(y)$ is determined by (7). Assumption MTR follows similarly, since under Assumption W(e),

$$\mathbb{P}\left[\Delta\mathring{\nu}(S_{it}(1)) - \Delta\mathring{\nu}(S_{it}(0)) \geq 0\right] = \mathbb{P}\left[W_{it}(1) \geq W_{it}(0)\right] = 1.$$

Q.E.D.

E Linearity of Parameters and Assumptions

The parameters and assumptions discussed in the main text can be represented as linear functions of $P = \{P(u, x) : u \in \mathcal{U}, x \in \mathcal{X}\}$. This section demonstrates this point. For notational simplicity, I assume throughout that X_i is degenerate, but it is straightforward to adjust the conditions to allow for X_i to be random by simply conditioning and then averaging over all realizations of X_i. (Alternatively, it is also straightforward to modify the parameters so that they are conditional on certain realizations of X_i.)

First, consider SD_t^+, which can be written as

$$\mathrm{SD}_t^+(P) \equiv \mathbb{P}_P[U_{it}(0) = 0, U_{it}(1) = 1] = \sum_{u \in \mathcal{U}_t^+} P(u),$$

where $\mathcal{U}_t^+$ is the set of $u = (u_0, u(0), u(1)) \in \mathcal{U}$ such that $u_t(0) = 0$ and $u_t(1) = 1$. This is a linear function of P. To see that $\mathrm{SD}_t^+(\cdot|0)$ is linear, write it as

$$\mathrm{SD}_t^+(P|0) = \frac{\mathbb{P}_P[U_{it}(0) = 0, U_{it}(1) = 1, Y_{it} = 0]}{\mathbb{P}[Y_{it} = 0]} = \frac{\sum_{u \in \mathcal{U}_t^+(0)} P(u)}{\mathbb{P}[Y_{it} = 0]},$$

where $\mathcal{U}_t^+(0)$ is the set of $u \in \mathcal{U}$ such that $u_t(0) = 0, u_t(1) = 1$, and $Y_{it} = 0$ when computed through the recursive relationship (1) with $Y_{i0} = u_0$, $U_{it}(0) = u_t(0)$ and $U_{it}(1) = u_t(1)$. Similar equations follow for $\mathrm{SD}_t^+(P|1), \mathrm{SD}_t^+(P|00)$ and $\mathrm{SD}_t^+(P|11)$.

To demonstrate linearity of the assumptions, consider Assumption MTR, which has the simplest form. Assumption MTR can be written as $\mathbb{P}_P[U_{it}(0) = 1, U_{it}(1) = 0] = 0$ for all $t \geq 1$. Hence, let $\mathcal{U}_t^{\mathrm{MTR}}$ denote the set of all $u \in \mathcal{U}$ such that $u_t(0) = 1$ and

$u_t(1) = 0$, and then write MTR as

$$\sum_{u \in \mathcal{U}_t^{\text{MTR}}} P(u) = 0, \tag{35}$$

for all $t \geq 1$. In terms of the ρ function, this equality constraint can be imposed with two inequalities.[36] Assumptions ST and its variations, TIV, MIV, and MATR can be imposed similarly by summing over the appropriate subsets of $\mathcal{U}$. Assumption MTS can be imposed using a construction similar to that for $\text{SD}_t^+(P|0)$.

Finally, consider Assumption DSC, which has a different structure. In general, Assumption DSC is a nonlinear restriction, but if Assumption ST holds so that the distribution of $U_{it}(d)$ does not depend on t, then $\text{Corr}_P(U_{it}(d), U_{i(t+s)}(d))$ is decreasing in $|s|$ if and only if $\text{Cov}_P(U_{it}(d), U_{i(t+s)}(d))$ is decreasing in $|s|$. Furthermore, under Assumption ST, the latter is true if and only if $\mathbb{E}_P[U_{it}(d)U_{i(t+s)}(d)]$, i.e. $\mathbb{P}_P[U_{it}(d) = 1, U_{i(t+s)}(d) = 1]$, is decreasing in $|s|$. It is straightforward to show that $\mathbb{P}_P[U_{it}(d) = 1, U_{i(t+s)}(d) = 1]$ is a linear function of P using a construction like (35).

F Dimension Reduction

F.1 Computational Considerations

The optimization problem in Proposition 2 can be quite large. For example, if $T = 6$, then the dimension of the variables in the problem, i.e. of $P = \{P(u, x) : u \in \mathcal{U}, x \in \mathcal{X}\}$ is $2^{2T+1} = 2^{13} = 8,192$, even without including any covariates. The number of constraints in the problem—even without any identifying assumptions—is at least $2^{T+1} = 128$ for the observational equivalence conditions (11), plus $2 \times 8{,}192$ constraints to ensure that P is contained in the unit interval.

These dimensions are large for an unstructured optimization problem. However, if both ρ and θ are linear so that the problems in Proposition 2 are linear programs, then these dimensions are actually fairly modest. A standard desktop computer with sophisticated linear programming solvers such as CPLEX (IBM, 2010) or Gurobi (Gurobi Optimization, 2015) or can finish problems of this size in a matter of seconds. Nevertheless, increasing the length of the panel, T, or including rich, time-varying specifications of covariates both increase the number of variables at an exponential rate. This can quickly become computationally infeasible.

In the remainder of this section, I describe three dimension reduction strategies for addressing this curse of dimensionality. The first strategy combines information across multiple models of shorter time horizons. The second strategy applies a similar construction to the covariates. The third strategy imposes a simple semiparametric structure for the covariates. All three strategies involve a natural and familiar trade-off between the amount of information in the data that is utilized, and the difficulty (both computational and statistical) of harnessing that information. They can be implemented separately or combined together.

[36] In practice, (35) is always combined with the requirement that $P(u) \geq 0$, and so it simply reduces to $P(u) = 0$ for all $u \in \mathcal{U}_t^{\text{MTR}}$.

F.2 Combining Shorter Models

The most immediate way in which the curse of dimensionality affects the DPO model is through the dimension of the potential outcomes sequence, U_i, which increases exponentially with the time period T. This difficulty is common for models that do not impose a conditional Markov restriction on the observed outcomes. For example, Hyslop (1999) considers a parametric DBR model in which the idiosyncratic error follows an AR(1) process. As observed by Heckman (1981a) and Chamberlain (1984), this implies that the observed outcomes are not Markov of any order. The resulting likelihood function for the parametric DBR involves a T–dimensional integral, which is also difficult to approximate when T is large.

In the DPO model, this strategy can be addressed by constructing several shorter, overlapping models. To see how this works, fix a *model length* $\text{ML} \in \{2, \ldots, T\}$. Then construct a DPO model for the observed sequence $(Y_{it_0}, Y_{i(t_0+1)}, \ldots, Y_{i(t_0+\text{ML})})$ at every initial period $t_0 \in \{0, 1, \ldots, T - \text{ML}\}$. Each of these shorter models relates potential outcomes to observed outcomes through (1) for $t \in \{t_0, \ldots, t_0 + \text{ML}\}$. The case discussed throughout the main text corresponds to setting $\text{ML} = T$.

The primitive is now a *collection* of probability mass functions $P \equiv \{P_{t_0}\}_{t_0=0}^{T-\text{ML}}$, each of which describes the joint distribution of

$$\left(Y_{it_0}, U_{i(t_0+1)}(0), \ldots, U_{i(t_0+\text{ML})}, U_{i(t_0+1)}(1), \ldots, U_{i(t_0+\text{ML})}(1), X_i\right).$$

Each P_{t_0} should satisfy (10), where $\mathcal{U}$ is now $\{0,1\}^{2\times\text{ML}+1}$, and each P_{t_0} is restricted to lie in a parameter space $\mathcal{P}_{t_0}^{\dagger}$, that can be specified to satisfy the same types of assumptions discussed in Section 4. The identified set contains all collections $P = \{P_{t_0} : P_{t_0} \in \mathcal{P}_{t_0}^{\dagger}\}_{t_0=0}^{T-\text{ML}}$ of shorter models such that the observational condition (11) is satisfied for each t_0, and all realizations of $(Y_{it_0}, Y_{i(t_0+1)}, \ldots, Y_{i(t_0+\text{ML})}, X_i)$. The target parameter, θ, is defined as a function P of the collection of shorter models.

Since P_{t_0} and P_{t_0+1} are distributions that encompass some of the same random variables, the identified set for P must satisfy an additional coherency condition. Mogstad, Torgovitsky, and Walters (2018b) describe such a condition (in a different model) as *logical consistency*. In the DPO model, the logical consistency condition states that when two overlapping models can both assign a probability to an event, this probability must be the same. That is,

$$\begin{aligned}\mathbb{P}_{P_{t_0}}\left[Y_{i(t_0+1)} = y_0, U_{i(t_0+2:t_0+\text{ML})} = u\right] &= \mathbb{P}_{P_{t_0+1}}\left[Y_{i(t_0+1)} = y_0, U_{i(t_0+2:t_0+\text{ML})} = u\right] \\ &\text{for all } (y_0, u) \in \{0,1\}^{1+2(\text{ML}-1)}. \end{aligned} \tag{36}$$

The identified set, $\mathcal{P}^{\star}$, only contains collections P that satisfy (36) for all t_0.[37] Intuitively, (36) aggregates information across the shorter overlapping models.

The benefit of this approach is that it reduces the dimension of variables of optimization from 2^{2T+1} to $(T - \text{ML})2^{2\times\text{ML}+1}$, which no longer increases exponentially with the length of the panel, T. However, it should also be noted that the coherency condition (36) constitutes an additional $(T - \text{ML})2^{2(\text{ML}-1)}$ constraints that are not

[37] Proposition 2 and the resulting methodology extend immediately, since these constraints are linear in each P_{t_0}.

present when ML $= T$. As a consequence, values of ML close to T may not provide any computational gain, and may in fact be costlier. For values of ML significantly smaller than T, however, the large reduction in the number of variables, combined with a modest increase in the number of constraints, can still net out to a massive dimension reduction.

The cost of this approach is the loss of information from modeling less of the distribution of Y_i.[38] This manifests itself in two ways. First, there are fewer observational equivalence conditions. This is because a model of $(Y_{it_0}, Y_{i(t_0+1)}, \ldots, Y_{i(t_0+\text{ML})})$ does not provide a probability for an observable sequence of length greater than ML $+ 1$. Second, it may not be possible to impose certain identifying assumptions, such as Assumption ST with $m >$ ML $- 1$, for the related reason that a model of length ML does not provide statements about potential outcome sequences longer than ML.

F.3 Partitioned Covariates

Recall that $\mathcal{X}$ denotes the support of the covariates X_i. For each $j = 1, \ldots, J$, let $\mathfrak{X}_j \equiv \{\mathcal{X}_{j1}, \ldots, \mathcal{X}_{jK_j}\}$ denote a finite partition of $\mathcal{X}$ into K_j exhaustive and mutually exclusive sets (or bins), $\mathcal{X}_{jk}$, that are specified by the researcher. Let $X_{ij} \equiv \sum_{k=1}^{K_j} k \mathbb{1}[X_i \in \mathcal{X}_{jk}]$ denote the random variable that takes value k if X_i lands in bin $\mathcal{X}_{jk}$.

Let P_j denote a probability mass function with support contained in $\mathcal{U} \times \mathfrak{X}_j$, and let $\mathcal{P}_j$ denote the set of all such functions that sum to unity, as in (10). Every $P \in \mathcal{P}$ defines a collection of $P_j \in \mathcal{P}_j$ for $j = 1, \ldots, J$ through the relationship

$$P_j : \mathcal{U} \times \mathbb{N} \to [0,1] : P_j(u,k) = \sum_{x \in \mathcal{X}_{jk}} P(u,x). \tag{37}$$

Given a parameter space, $\mathcal{P}^\dagger$, relationship (37) generates parameter spaces $\mathcal{P}_j^\dagger$ that each P_j is restricted to lie in. Moreover, if $P \in \mathcal{P}^\star$, then P_j must satisfy the following set of observational equivalence constraints:

$$\begin{aligned} \mathbb{P}[Y_i = y, X_i \in \mathcal{X}_{jk}] &= \sum_{x \in \mathcal{X}_{jk}} \mathbb{P}[Y_i = y, X_i = x] \\ &= \sum_{x \in \mathcal{X}_{jk}} \sum_{u \in \mathcal{U}_{\text{oeq}}(y)} P(u,x) = \sum_{u \in \mathcal{U}_{\text{oeq}}(y)} P_j(u,k), \text{ for all } y, j \text{ and } k. \end{aligned} \tag{38}$$

Similar to the previous section, there is an additional logical consistency constraint that can be imposed across the smaller models $\{P_j\}_{j=1}^J$, namely:

$$\sum_{k=1}^{K_j} P_j(u,k) = \sum_{x \in \mathcal{X}} P(u,x) = \sum_{k=1}^{K_{j'}} P_{j'}(u,k) \text{ for any } j, j' \text{ and all } u \in \mathcal{U}. \tag{39}$$

As long as the collection $\{P_j\}_{j=1}^J$ is sufficient for evaluating the researcher's target parameter, θ, one can work only with these lower dimensional objects via (37)–(39), rather than with P directly.

[38] Formally, the identified set will be an outer identified set, rather than the sharp identified set.

To see how this partitioning approach addresses the curse of dimensionality, suppose that $X_i = (X_{i1}, \ldots, X_{id_x})$ and that each X_{ij} has K support points $\{x_{jk}\}_{k=1}^K$. The full distribution P consists of $2^{2T+1} \times J^K$ elements. This can be a large number even if J or K are relatively small. However, suppose that the researcher specifies partitions $\mathfrak{X}_j$ taken as $\mathfrak{X}_j = \{\{x_{1k}\}_{k=1}^K, \ldots, \{x_{Jk}\}_{k=1}^K\}$, so that each partition has K elements corresponding to the jth component of X_i. The total dimension of $\{P_j\}_{j=1}^J$ under this partition is $2^{2T+1} \times JK$, which can be dramatically smaller than J^K. The cost of this approach is a loss of information. This occurs for the same reasons as for the strategy in the previous section: Only a subset of the observational equivalence conditions are being met, and identifying content contained in restrictions that would span across covariate partitions cannot be exploited.

F.4 A Semiparametric Specification

One natural response to the curse of dimensionality is to impose semiparametric restrictions. For doing this, it is more convenient to formulate the DPO model as one of the *conditional* distribution of Y_i given X_i, rather than of their joint distribution. This changes most of the discussion in the paper in only obvious ways; the exception is the discussion of statistical inference in Appendix G, which would require some reworking. The primitive object of the model changes from a joint distribution to a collection of conditional distributions, written (with mild abuse of notation) as $P = \{P(\cdot|x) : x \in \mathcal{X}\}$, each of which has support contained in $\mathcal{U} \equiv \{0,1\}^{2T+1}$.

With P as a conditional distribution, a natural semiparametric assumption is that

$$\mathcal{P}^\dagger \subseteq \left\{ P \in \mathcal{P} : P(u|x) = h(x)'\beta_u \text{ for some } \beta_u \in \mathbb{R}^{d_h}, \text{ all } u \text{ and } x \right\},$$

where h is a known, vector-valued function of length d_h. This assumption says that for each u, $P(u|x)$ is a linear function of a known transformation of x with coefficient vector β_u.[39] Under this assumption, each P is characterized by a set of parameters $\{\beta_u : u \in \mathcal{U}\}$ that has dimension $2^{2T+1} \times d_h$. This dimension does not depend on the number of support points of X_i, and grows linearly with the dimension of the transformation vector, h, thereby overcoming the curse of dimensionality, while preserving the linear programming structure. The cost is the usual threat of potential misspecification.

When taken to the observational equivalence condition (11), the semiparametric model also implies a lower-dimensional representation for the observed data distribution, since

$$\mathbb{P}_P[Y_i = y|X_i = x] = \sum_{u \in \mathcal{U}_{\text{oeq}}(y)} P(u|x) = h(x)' \left(\sum_{u \in \mathcal{U}_{\text{oeq}}(y)} \beta_u \right) \equiv h(x)'\delta_y. \tag{40}$$

Thus, it justifies estimating $\mathbb{P}[Y_i = y|X_i = x]$ by a linear probability model. In practice, one would want to ensure that the fitted probabilities are in the unit interval. One

[39] Note that unlike a linear probability model, here β_u is still required to be such that $P(u|x) \in [0,1]$.

way to do this is to estimate a constrained least squares regression

$$\hat{\delta}_y \equiv \underset{\delta \in \mathbb{R}^{d_h}}{\arg\min} \sum_{i=1}^{n} (\mathbb{1}[Y_i = y] - h(X_i)'\delta)^2 \quad \text{s.t.} \quad 0 \leq h(X_i)'\delta \leq 1 \; \forall i = 1, \ldots, n, \quad (41)$$

as suggested by Domencich and McFadden (1975, pg. 105) or Judge, Griffiths, Hill, Lütkepohl, and Lee (1985, pg. 759). Alternatively, one can estimate $\mathbb{P}[Y_i = y | X_i = x]$ using a statistical model chosen for fit, and then run the resulting probabilities through the DPO model.

G Estimation and Statistical Inference

In Section 3, I considered the distribution of observables as if it were known without accounting for any sampling error. As a result, the identified set $\Theta^\star$ was also known without error for a given parameter and given set of assumptions. In this section, I adjust the discussion to account for the statistical variation that arises when modeling the data as an i.i.d. sample from some underlying population distribution. First, I describe how to construct a consistent estimator of $\Theta^\star$. Second, I describe how to construct confidence regions that contain (with probability at least $1-\alpha$) the parameter $\theta_0 = \theta(P_0) \in \Theta^\star$ corresponding to the "true" $P_0 \in \mathcal{P}^\star$ that generated the data. Third, I discuss a specification test that can be used to falsify the hypothesis that such a P_0 exists. Finally, I conduct a Monte Carlo simulation to evaluate these procedures.

G.1 The Criterion Function

Approaches based on direct sample analogs of $\theta^\star_{\text{lb}}$ and $\theta^\star_{\text{ub}}$ are unattractive for two important reasons. First, while these estimators are consistent under weak conditions, their asymptotic distributions are highly nonstandard.[40] Second, sample analogs of $\theta^\star_{\text{lb}}$ and $\theta^\star_{\text{ub}}$ might not exist even when the population identified set is non-empty.[41] Both problems can be addressed by transforming the characterization of the identified set provided in Proposition 2 into a criterion function, and then using a sample analog of this criterion function as the basis for estimation and statistical inference (e.g. Chernozhukov, Hong, and Tamer, 2007).

Some additional notation is required. Let $\mathcal{W} \equiv \text{supp}(Y_i, X_i)$ denote the joint support of the observable data $W_i \equiv (Y_i, X_i)$. For each $w \equiv (w_y, w_x) \in \mathcal{W} \subset \mathbb{R}^{d_W}$ define

$$m_{\text{oeq},w}(W_i, P) \equiv \mathbb{1}[Y_i = w_y, X_i = w_x] - \sum_{u \in \mathcal{U}_{\text{oeq}}(w_y)} P(u, w_x). \quad (42)$$

[40] See Shapiro and Dentcheva (2014, Chapter 5), who derive the asymptotic distributions of these analog estimators. The results of Andrews and Han (2009) imply that naively bootstrapping or subsampling empirical analogs of $\theta^\star_{\text{lb}}$ and $\theta^\star_{\text{ub}}$ will not lead to valid confidence regions.

[41] Freyberger and Horowitz (2015) study an instrumental variables model for which the identified set can be represented through the solution to two linear programming problems. They propose a modified bootstrap procedure based on the sample analogs of the solutions to the linear programs, but their procedure assumes that these sample analogs exist.

Next, partition the restriction function ρ into a deterministic component ρ_d, and a stochastic component ρ_s with dimension d_s. The deterministic component, $\rho_d : \mathcal{P} \to \mathbb{R}^{d_\rho - d_s}$, is a function defined on $\mathcal{P}$ that does not depend on the distribution of W_i. The stochastic component, $\rho_s : \mathcal{P} \to \mathbb{R}^{d_s}$, is a function defined on $\mathcal{P}$ that is assumed to be representable as a moment condition. That is, it is assumed that there exists a function $m_\rho : \mathcal{W} \times \mathcal{P} \to \mathbb{R}^{d_s}$ for which $\rho_s(P) = \mathbb{E}\, m_\rho(W_i, P)$. This condition is satisfied by all of the identifying assumptions discussed in Section 4.

For example, Assumption ST would be part of ρ_d, since it is not a restriction that depends on the distribution of observables W_i. On the other hand, Assumption MTS would be part of ρ_s, since it depends on the distribution of $(Y_{i(t-1)}, Y_{i(t-2)})$.

Next, define $\mathcal{P}_d^\dagger \equiv \{P \in \mathcal{P} : \rho_d(P) \geq 0\}$ as the set of deterministic constraints on P. These include not only ρ_d, but also the requirement that $P \in \mathcal{P}$, i.e. that P is a probability mass function on $\mathcal{U} \times \mathcal{X}$. Then

$$\begin{aligned} \mathcal{P}^\star = \{P \in \mathcal{P}_d^\dagger : \mathbb{E}\, m_{\text{oeq},w}(W_i, P) = 0 \; \forall w \in \mathcal{W} \\ \text{and } \mathbb{E}\, m_{\rho,s}(W_i, P) \geq 0 \; \forall s = 1, \ldots, d_s\}, \end{aligned} \tag{43}$$

where $m_{\rho,s}(W_i, P)$ denotes the s^{th} component of $m_\rho(W_i, P)$. Equation (43) shows that the DPO model can be viewed as a moment inequality model with parameter space $\mathcal{P}_d^\dagger$, moment equalities $\{\mathbb{E}\, m_{\text{oeq},w}(W_i, P) = 0\}_{w \in \mathcal{W}}$, and moment inequalities $\{\mathbb{E}\, m_{\rho,s}(W_i, P) \geq 0\}_{s=1}^{d_s}$. Alternatively and equivalently, let $\eta \in \mathbb{R}_+^{d_s}$ denote a vector of non-negative slackness variables and define the identified set using only moment equalities as

$$\begin{aligned} \mathcal{R}^\star \equiv \{(P, \eta) \in \mathcal{P}_d^\dagger \times \mathbb{R}_+^{d_s} : \mathbb{E}\, m_{\text{oeq},w}(W_i, P) = 0 \; \forall w \in \mathcal{W} \\ \text{and } \mathbb{E}\, m_{\rho,s}(W_i, P) - \eta_s = 0 \; \forall s = 1, \ldots, d_s\}. \end{aligned} \tag{44}$$

Then $\mathcal{P}^\star$ is the projection of the first component of $\mathcal{R}^\star$, i.e.

$$\mathcal{P}^\star = \{P \in \mathcal{P} : (P, \eta) \in \mathcal{R}^\star \text{ for some } \eta \in \mathbb{R}_+^{d_s}\}.$$

Write the moment functions $\{m_{\rho,s}\}_{s=1}^{d_s}$ and $\{m_{\text{oeq},w}\}_{w \in \mathcal{W}}$ together as $\{m_j\}_{j=1}^{d_m}$ where $d_m = d_s + d_W$ and the first d_s components of $\{m_j\}_{j=1}^{d_m}$ correspond to $\{m_{\rho,s}\}_{s=1}^{d_s}$. A convenient choice of population criterion function is

$$Q(P, \eta) \equiv \sum_{j=1}^{d_s} |\mathbb{E}\, m_j(W_i, P) - \eta_j| + \sum_{j=d_s+1}^{d_m} |\mathbb{E}\, m_j(W_i, P)|. \tag{45}$$

Notice that $(P, \eta) \in \mathcal{R}^\star$ if and only if $Q(P, \eta) = 0$ and $(P, \eta) \in \mathcal{P}_d^\dagger \times \mathbb{R}_+^{d_s}$. Using an absolute value loss function, instead of the more standard quadratic loss function, is computationally convenient for the estimation and inference procedures discussed ahead. Other choices of criterion function are possible in principle, but tend to create computational obstacles.[42] Given an i.i.d. sample $\{W_i\}_{i=1}^n$ of size n, a sample analog

[42] In a previous draft of this paper, I used a quadratic criterion function. The finite sample performance

of Q is constructed by replacing the population expectation with its (scaled) empirical counterpart:

$$Q_n(P,\eta) \equiv \sum_{j=1}^{d_s} \sqrt{n}\,|\bar{m}_{n,j}(P) - \eta_s| + \sum_{j=d_s+1}^{d_m} \sqrt{n}\,|\bar{m}_{n,j}(P)|\,, \tag{46}$$

$$\text{where } \bar{m}_{n,j}(P) \equiv \frac{1}{n}\sum_{i=1}^{n} m_j(W_i,P) \text{ for } j = 1,\ldots,d_m.$$

G.2 Estimation

An estimator of $\Theta^\star$ can be constructed by restricting attention to P that come close to minimizing the sample criterion (46). Let

$$\bar{Q}_n \equiv \min_{(P,\eta)\in\mathcal{P}_d^\dagger\times\mathbb{R}_+^{d_s}} Q_n(P,\eta) \tag{47}$$

denote the minimum value of Q_n, and let

$$\mathcal{P}_n \equiv \left\{P \in \mathcal{P}_d^\dagger : Q_n(P,\eta) \leq \bar{Q}_n(1+\tau_n) \text{ for some } \eta \in \mathbb{R}_+^{d_s}\right\}$$

denote the collection of $P \in \mathcal{P}_d^\dagger$ that yield criterion values within $\tau_n\%$ of the optimum. Then define

$$\hat{\theta}^\star_{\text{lb}} \equiv \min\theta(\mathcal{P}_n) = \min_{(P,\eta)\in\mathcal{P}_d^\dagger\times\mathbb{R}_+^{d_s}} \theta(P) \text{ s.t. } Q_n(P,\eta) \leq \bar{Q}_n(1+\tau_n)$$

$$\text{and} \quad \hat{\theta}^\star_{\text{ub}} \equiv \max\theta(\mathcal{P}_n) = \max_{(P,\eta)\in\mathcal{P}_d^\dagger\times\mathbb{R}_+^{d_s}} \theta(P) \text{ s.t. } Q_n(P,\eta) \leq \bar{Q}_n(1+\tau_n).$$

Theorem S.1 of Mogstad, Santos, and Torgovitsky (2018a) shows that $[\hat{\theta}^\star_{\text{lb}}, \hat{\theta}^\star_{\text{ub}}]$ will be a consistent estimator of $\Theta^\star$ in the Hausdorff metric. The result requires $\Theta^\star$ to be non-empty, so that the model is correctly specified. It also requires $\tau_n \to 0$. For the empirical results in Section 5, I used $\tau_n = .25$. This value was chosen because confidence regions constructed using the method discussed in the next section require a similar tuning parameter, and $\tau_n = .25$ performed well in terms of size for the Monte Carlo simulations discussed in Appendix G.6.

was slightly better, but the computation was significantly more difficult for reasons I discuss further in Appendix G.5. Also, criterion functions that incorporate information on the covariance matrix for the moments may have preferable statistical properties; see AS and Andrews and Barwick (2012). However, Studentizing the moments introduces non-linearities when the moments are not additively separable in P. This severely complicates computation, because the constraint sets for the optimization problems proposed ahead become potentially non-convex.

G.3 Confidence Regions

As is common in the literature on inference under partial identification, I will construct confidence regions through test inversion.[43] The tests will be of null hypotheses taking the form $H_0 : t \in \Theta^\star$ for conjectured scalar values t. A natural test statistic for this null is a profiled version of Q_n:

$$\bar{Q}_n(t) \equiv \inf_{(P,\eta)\in\mathcal{P}_d^\dagger(t)\times\mathbb{R}_+^{d_s}} Q_n(P,\eta), \tag{48}$$

where $\mathcal{P}_d^\dagger(t) \equiv \{P \in \mathcal{P}_d^\dagger : \theta(P) = t\}$. Constructing a confidence region for $\Theta^\star$ by inverting these tests means collecting all t for which $\bar{Q}_n(t)$ is not "too large."

To operationalize such a test, one needs to determine how large is "too large" by approximating the distribution of $\bar{Q}_n(t)$ under the null hypothesis. The asymptotic distribution of $Q_n(P,\eta)$ is itself non-standard due to the lack of point identification, see for example Chernozhukov et al. (2007), Andrews and Soares (2010), Bugni (2010) and Canay (2010). There is an added difficulty here caused by the infimum in the definition of $\bar{Q}_n(t)$, which is introduced by the desire to conduct profile (or "subvector") inference on $\Theta^\star$ rather than $\mathcal{P}^\star$. The two procedures I consider for this problem are subsampling and the shape restriction approach of Chernozhukov et al. (2015).[44]

The subsampling approach approximates the distribution of $\bar{Q}_n(t)$ under the null hypothesis with the distribution of

$$\bar{Q}_b^{\mathrm{SS}}(t) \equiv \inf_{(P,\eta)\in\mathcal{P}_d^\dagger(t)\times\mathbb{R}_+^{d_s}} Q_b^{\mathrm{SS}}(P,\eta),$$

where $Q_b^{\mathrm{SS}}(P,\eta)$ is defined analogously to $Q_n(P,\eta)$, but constructed instead using a subsample $\{W_i^\star\}_{i=1}^b$ of size b that is randomly drawn (without replacement) from $\{W_i\}_{i=1}^n$. This profiled subsampling procedure was first proposed in Romano and Shaikh (2008), see also Chernozhukov et al. (2007) and Romano and Shaikh (2010). In the following, I refer to the test that rejects $H_0 : t \in \Theta^\star$ when $\bar{Q}_n(t)$ is greater than the $1-\alpha$ quantile of $\bar{Q}_b^{\mathrm{SS}}(t)$ based on B random subsamples as the SS test. A $1-\alpha$ SS confidence region for $\Theta^\star$ is the set of all t for which the SS test does not reject.

The Monte Carlo simulations in the next section suggest that the SS test can be quite conservative in the DPO model. This leads to low power and excessively wide confidence regions. The procedure for testing shape constraints recently proposed by Chernozhukov et al. (2015, "CNS") provides an alternative that turns out to be less conservative in the DPO model. Their approach is based on a careful approximation of $\bar{Q}_n(t)$ that takes into account the shape of the constraint set $\mathcal{P}_d^\dagger(t) \times \mathbb{R}_+^{d_s}$.

[43] See Canay and Shaikh (2017) for a recent survey of the literature.

[44] In a previous version of this paper, I also applied the method proposed by Bugni, Canay, and Shi (2017). Monte Carlo results reported in that version of the paper suggest that this approach has low power in the DPO model. Another recently proposed procedure for profile inference in partially identified models is Kaido, Molinari, and Stoye (2016). Unfortunately, their approach is not computationally feasible for the dimension of the nuisance parameters (P) considered here.

To be more specific, first define the function

$$Q_n^\star(P,\eta,g,h) \equiv \sum_{j=1}^{d_s} \left| \xi_{n,j}^\star(P) + \frac{1}{n}\sum_{i=1}^n \nabla m_j(W_i,P)[g] - h_j \right| + \sum_{j=d_s+1}^{d_m} \left| \xi_{n,j}^\star(P) + \frac{1}{n}\sum_{i=1}^n \nabla m_j(W_i,P)[g] \right|.$$

Here, (g,h) are parameters that serve as local deviations to (P,η) and, correspondingly, have dimensions 2^{2T+1} and d_s, respectively. The notation $\nabla m_j(W_i,P)[g]$ stands for the directional derivative of m_j with respect to P, evaluated at P, in the direction g, i.e. $\frac{\partial}{\partial\kappa} m_j(W_i, P+\kappa g)|_{\kappa=0}$. The function $\xi_{n,j}^\star$ is defined for each $j = 1,\ldots,d_m$ as

$$\xi_{n,j}^\star(P) \equiv \frac{1}{\sqrt{n}}\sum_{i=1}^n \left[m_j(W_i^\star, P) - \bar{m}_{n,j}(P) \right],$$

where $\{W_i^\star\}_{i=1}^n$ is a bootstrap sample drawn i.i.d. with replacement from $\{W_i\}_{i=1}^n$. CNS then approximate the distribution of $\bar{Q}_n(t)$ with that of

$$\begin{aligned} \tilde{Q}_n(t) \equiv \min_{(P,\eta,g,h)} & Q_n^\star(P,\eta,g,h) \\ & \text{s.t. } (P,\eta) \in \widehat{\mathcal{R}}^\star(t) \\ & \text{and } (P,\eta) + n^{-1/2}(g,h) \in \mathcal{P}_d^\dagger(t) \times \mathbb{R}_+^{d_s}. \end{aligned} \tag{49}$$

The second constraint here restricts (g,h) to be small deviations of (P,η) that remain inside the admissible parameter space under the null hypothesis.[45] The first constraint uses the definition

$$\widehat{\mathcal{R}}^\star(t) \equiv \left\{ (P,\eta) \in \mathcal{P}_d^\dagger(t) \times \mathbb{R}_+^{d_s} : Q_n(P,\eta) \le \bar{Q}_n(t)(1+\tau_n) \right\},$$

which is the set of (P,η) that approximately solve (48). In the Monte Carlo simulations in the next section I find that $\tau_n = .25$ works well in terms of size, so this is the value that I use in the application in Section 5.

The distribution of $\tilde{Q}_n(t)$ can be approximated by redrawing $\{W_i^\star\}_{i=1}^n$ a large number (B) of times and computing $\tilde{Q}_n(t)$ for each draw. I refer to the test that rejects $H_0 : t \in \Theta^\star$ when $\bar{Q}_n(t)$ is greater than the $1-\alpha$ quantile of these B values of $\tilde{Q}_n(t)$ as the CNS test. A $1-\alpha$ CNS confidence region for $\Theta^\star$ is the set of all t for which the CNS test does not reject.

CNS provide a set of sufficient conditions under which their procedure controls size uniformly (see their Theorem 6.3). Most of the sufficient conditions are satisfied immediately here because $\mathcal{P}$ is finite-dimensional and compact, and both the moment

[45] CNS include an additional tuning parameter that regulates the slackness of inequality constraints in $\mathcal{P}_d^\dagger(t) \times \mathbb{R}_+^{d_s}$. This parameter is theoretically necessary, however it introduces a non-convexity into (49) that renders the problem computationally intractable. The Monte Carlo simulations in Appendix G.6 suggest the CNS test performs well in the DPO model even when this parameter is omitted.

conditions and constraints on $\mathcal{P}$ are linear in P. This includes CNS's Assumptions 2.1–2.2, 3.2–3.3, 4.1–4.2, 5.1–5.4, 6.1–6.5. Their Assumption 3.1 is satisfied when the sample is drawn i.i.d., and their Assumption 3.4 is satisfied because the objective function is unweighted. Assumption 6.6 is a rate condition on the tuning parameter τ_n and the additional tuning parameter used by CNS that I am excluding for computational considerations; see footnote 45. The sufficient conditions also include a high-level anti-concentration condition (their Assumption 6.7), which is mild in a finite-dimensional setting.

G.4 Testing for Misspecification

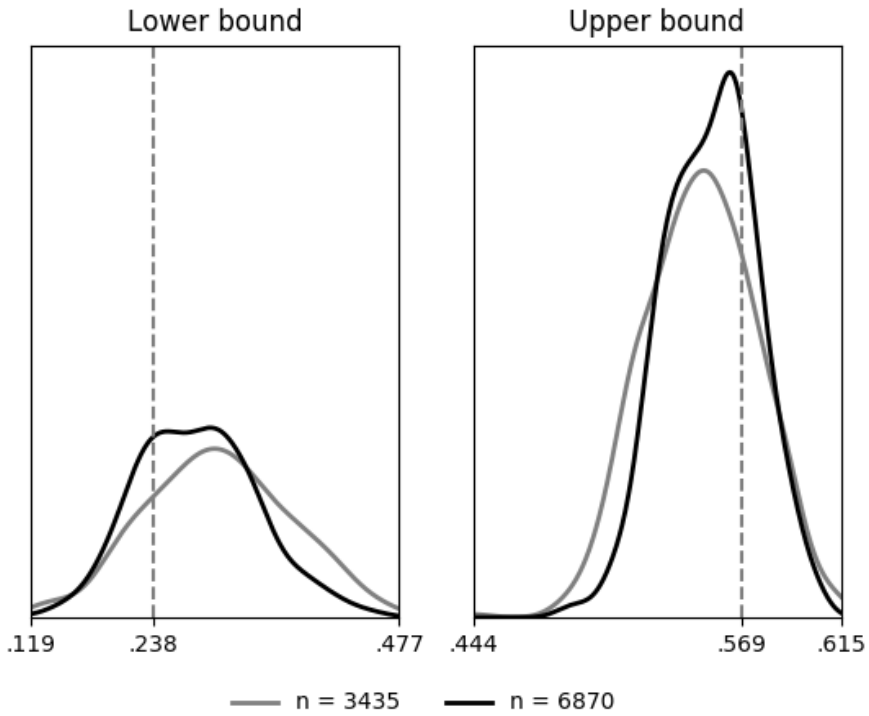
An attractive feature of a partial identification strategy is that it provides an immediate specification test based on the nonemptyness of the identified set. Specifically, a rejection of the null hypothesis $H_0 : \mathcal{P}^\star \neq \emptyset$ is evidence of the nonexistence of an admissible $P \in \mathcal{P}^\dagger$ that is consistent with the data, and hence evidence that some of the assumptions embodied in $\mathcal{P}^\dagger$ are false, i.e. that the model is misspecified. A natural statistic for such a test is the overall minimum criterion value, $\bar{Q}_n$, defined in (47). The results of CNS show that one can approximate the distribution of $\bar{Q}_n$ by simulating the distribution of a quantity that is analogous to (49), but which replaces $\mathcal{P}_d^\dagger(t)$ by $\mathcal{P}_d^\dagger$ throughout. A level α misspecification test rejects the null that the identified set is non-empty when $\bar{Q}_n$ is greater than the $1-\alpha$ quantile of this simulated distribution. Note that such a test always fails to reject when the estimated identified set is non-empty, since in such cases $\bar{Q}_n = 0$.

G.5 Computing Critical Values

In order to implement the SS and CNS tests, it is important to be able to reliably solve the optimization problems that define $\bar{Q}_n(t)$ and $\tilde{Q}_n(t)$. Reliability—in particular, ensuring that local optima are in fact global optima—is especially important here because each problem needs to be solved a large number of times in the process of resampling and inverting hypothesis tests to construct confidence regions. Both problems are convex as long as each $\bar{m}_{n,j}(P)$ is linear in P for every j and $\mathcal{P}_d^\dagger(t)$ is determined by the intersection of linear equalities and inequalities. These conditions are satisfied by all of the assumptions and parameters discussed in Section 2.

Under these conditions, the optimization problem in the definition of $\bar{Q}_n(t)$ (and hence $\bar{Q}_b^{\text{SS}}(t)$) can be reformulated as a linear program, using a standard reformulation argument for the absolute value function. As a result, solving this problem is not significantly harder than solving the linear programs used to directly estimate the bounds of the identified set. The optimization problem in the definition of $\tilde{Q}_n(t)$ can also be shown to be a linear program, again by reformulating the absolute value function. This is the motivation for choosing the absolute loss function in (45) rather than a quadratic loss function. With quadratic loss, (49) would be a quadratically-constrained quadratic program, due to the definition of $\widehat{\mathcal{R}}^\star(t)$. While such programs are still convex, and can still be solved reliably using widely-available solvers, they are significantly more costly to solve than the corresponding problem using an absolute

Figure G.1: Density of Estimated Bounds in Column (6) of Table 2



Notes: The plot shows kernel density estimates of the directly estimated lower and upper bounds of $\text{SD}^{+}_{\text{avg}}(\cdot|0)$ from the Monte Carlo reported in Table G.1. The gray dotted lines indicate the true lower and upper bounds of the identified set in the DGP. The smoothing used a Gaussian kernel and Silverman's rule-of-thumb bandwidth.

loss function.[46]

G.6 Monte Carlo Simulations

In this section I report the results of a Monte Carlo study that evaluates the procedures discussed in the preceding sections. The data generating process in the study draws Y_i according to the empirical probabilities in the SIPP sample used in Section 5.

Table G.1 reports the finite-sample properties of the estimated bounds, $\hat{\theta}^{\star}_{\text{lb}}$ and $\hat{\theta}^{\star}_{\text{ub}}$, using the same time horizon as in the application ($T = 6$), and maintaining Assumption ST with $m = 4$ as in column (6) of Table 2. The statistics are based on 500 simulation draws, and the tuning parameter τ_n is set to .25. Comparing results across increasing sample sizes suggests that the bound estimators are consistent. Figure G.1 plots the estimated density of $\hat{\theta}^{\star}_{\text{lb}}$ and $\hat{\theta}^{\star}_{\text{ub}}$ when the target parameter is $\text{SD}^{+}_{\text{avg}}(\cdot|0)$.

[46] In previous drafts of this paper, I used a quadratic loss function and solved the quadratically-constrained quadratic programs. This procedure was substantially more computationally demanding.

The distributions are non-normal and non-standard, which accords with theoretical predictions, see e.g. Section 5 of Shapiro and Dentcheva (2014).

Table G.2 reports the rejection rates of the SS and CNS tests of $H_0 : t \in \Theta^\star$ at nominal levels $\alpha = .01$, .05, and .10 for several values of t. The data generating process is still the SIPP sample, but I only use the first four periods ($T = 3$) in order to moderate the computational burden. The maintained assumptions are Assumption ST with $m = 1$ and Assumption MTR, with the latter assumption also being useful for easing computation. The reported statistics are based on 500 simulation draws, 500 bootstrap draws or subsamples per simulation, and still uses the tuning parameter $\tau_n = .25$. The sample size is set at $n = 3,435$, as in the application, and the subsample size is set to $n^{2/3} \approx 228$. The target parameter is taken to be $\mathrm{SD}^+_{\mathrm{avg}}(\cdot|0)$. For comparison, Table G.3 reports the finite-sample performance for the estimated bounds in this case.

The results for t at the boundary of the identified set suggest that the SS test is quite conservative. This leads to low power when testing points outside of the identified set, and thus large confidence regions when constructing these regions through test inversion. In contrast, the CNS test maintains roughly the nominal level at the boundary of the identified set, and is much more powerful at points outside of the identified set. Table G.3 suggests that the CNS test may still have low power in this setting; for example it rejects $t = .150$ only about 72% of the time, even though this point is more than three standard deviations smaller than the lower bound of the identified set. These results provide reassurance that the CNS test works reasonably well for the DPO model, at least for the empirical setting considered here, and provides at least a conservative indication of the impacts of statistical uncertainty.

Table G.1: Finite Sample Properties of Estimated Bounds in Column (6) of Table 2

		$\hat{\theta}^{\star}_{\text{lb}}$			$\hat{\theta}^{\star}_{\text{ub}}$		
	sample size	1718	3435	6870	1718	3435	6870
$\text{SD}^{+}_{\text{avg}}$	true	.034	.034	.034	.933	.933	.933
	mean	.047	.043	.040	.931	.931	.932
	std	.009	.009	.007	.006	.004	.003
	rmse	.015	.013	.010	.006	.005	.003
	5/95%	.031	.029	.030	.940	.938	.937
	min/max	.017	.020	.020	.947	.942	.941
$\text{SD}^{+}_{\text{avg}}(\cdot\|0)$	true	.238	.238	.238	.569	.569	.569
	mean	.319	.300	.281	.546	.550	.554
	std	.067	.070	.058	.029	.026	.021
	rmse	.105	.093	.072	.037	.032	.026
	5/95%	.207	.193	.193	.594	.591	.589
	min/max	.120	.123	.119	.622	.615	.605
$\text{SD}^{+}_{\text{avg}}(\cdot\|00)$	true	.414	.414	.414	.980	.980	.980
	mean	.553	.521	.487	.942	.948	.956
	std	.119	.125	.104	.032	.030	.026
	rmse	.183	.165	.127	.050	.044	.035
	5/95%	.348	.329	.328	.988	.993	.994
	min/max	.193	.214	.202	1.00	1.00	1.00
$\text{SD}^{+}_{\text{avg}}(\cdot\|1)$	true	.016	.016	.016	.963	.963	.963
	mean	.026	.024	.021	.959	.960	.961
	std	.007	.006	.005	.005	.004	.003
	rmse	.012	.010	.007	.006	.005	.004
	5/95%	.016	.014	.014	.966	.966	.966
	min/max	.006	.009	.011	.972	.969	.968
$\text{SD}^{+}_{\text{avg}}(\cdot\|11)$	true	.017	.017	.017	.993	.993	.993
	mean	.027	.024	.022	.989	.990	.991
	std	.007	.006	.005	.004	.004	.003
	rmse	.012	.010	.007	.006	.005	.003
	5/95%	.016	.015	.014	.995	.995	.995
	min/max	.006	.009	.011	.998	.997	.997
$\mathbb{P}[\Theta^{\star} = \emptyset$ in sample]		.882	.624	.334	–	–	–

Notes: The bounds are computed with $T = 6$ under Assumption ST with $m = 4$. The row 5/95% gives the .05 quantile across simulations of $\hat{\theta}^{\star}_{\text{lb}}$ and the .95 quantile of $\hat{\theta}^{\star}_{\text{ub}}$. Similarly, the row min/max gives the minimum across simulations of $\hat{\theta}^{\star}_{\text{lb}}$ and the maximum across simulations of $\hat{\theta}^{\star}_{\text{ub}}$. The final row shows the proportion of simulations in which the sample identified set is empty. The statistics are based on 500 replications and the tuning parameter is set at $\tau_n = .25$.

Table G.2: Finite Sample Rejection Probabilities of SS and CNS Tests

		rejection probability of $H_0 : t \in \Theta^\star$ for $t = \ldots$											
level	test	.090	.150	.210	.270	.300	.338	.445	.480	.510	.540	.600	.660
.01	SS	.098	.032	.004	.000	.000	.000	.000	.002	.004	.018	.140	.578
	CNS	.724	.478	.242	.078	.034	.012	.008	.042	.146	.290	.724	.982
.05	SS	.408	.190	.066	.012	.006	.004	.006	.014	.036	.094	.394	.912
	CNS	.876	.716	.472	.204	.118	.048	.050	.158	.326	.522	.906	1.00
.10	SS	.602	.394	.164	.046	.016	.010	.010	.032	.094	.208	.588	.984
	CNS	.920	.802	.596	.330	.200	.092	.102	.242	.430	.648	.964	1.00

Notes: The target parameter is $\mathrm{SD}^{+}_{\mathrm{avg}}(\cdot|0)$ and Assumptions ST (with $m = 1$) and MTR are maintained. The time horizon is $T = 3$. The values of t in boxes indicate the lower and upper bound of the population identified set. The statistics are based on 500 replications, 500 bootstraps (for CNS) or subsamples (for SS, with $b = n^{2/3} = 228$), and the tuning parameter is set at $\tau_n = .25$.

Table G.3: Finite Sample Properties of Estimated Bounds from Table G.2

		$\hat{\theta}^{\star}_{\text{lb}}$			$\hat{\theta}^{\star}_{\text{ub}}$		
	sample size	1718	3435	6870	1718	3435	6870
$\text{SD}^{+}_{\text{avg}}$	true	.037	.037	.037	.925	.925	.925
	mean	.030	.032	.034	.925	.925	.925
	std	.009	.007	.005	.006	.005	.003
	rmse	.011	.008	.006	.006	.005	.003
	5/95%	.014	.019	.026	.934	.932	.931
	min/max	.001	.010	.017	.942	.938	.935
$\text{SD}^{+}_{\text{avg}}(\cdot\|0)$	true	.338	.338	.338	.445	.445	.445
	mean	.274	.295	.316	.457	.454	.448
	std	.075	.062	.045	.040	.036	.028
	rmse	.099	.075	.050	.042	.037	.028
	5/95%	.126	.172	.235	.527	.520	.499
	min/max	.006	.077	.168	.595	.587	.548
$\text{SD}^{+}_{\text{avg}}(\cdot\|00)$	true	.614	.614	.614	.807	.807	.807
	mean	.500	.537	.575	.828	.821	.812
	std	.136	.115	.084	.057	.051	.041
	rmse	.177	.138	.093	.060	.053	.041
	5/95%	.221	.307	.422	.927	.914	.882
	min/max	.011	.141	.286	.962	.989	.930
$\text{SD}^{+}_{\text{avg}}(\cdot\|1)$	true	.016	.016	.016	.957	.957	.957
	mean	.013	.014	.015	.957	.957	.957
	std	.004	.003	.003	.005	.004	.003
	rmse	.005	.004	.003	.005	.004	.003
	5/95%	.006	.008	.011	.964	.962	.961
	min/max	.000	.004	.006	.968	.967	.965
$\text{SD}^{+}_{\text{avg}}(\cdot\|11)$	true	.017	.017	.017	.990	.990	.990
	mean	.014	.015	.016	.990	.990	.990
	std	.004	.004	.003	.003	.003	.002
	rmse	.005	.004	.003	.003	.003	.002
	5/95%	.006	.008	.011	.995	.994	.993
	min/max	.000	.004	.007	.998	.996	.996
$\mathbb{P}[\Theta^{\star} = \emptyset$ in sample]		.428	.358	.290	–	–	–

Notes: The bounds are computed with $T = 3$ under Assumption ST with $m = 1$ and Assumption MTR. See notes for Table G.1.

H Additional Empirical Estimates

Table H.1: Additional Estimates

	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**
				Assumptions				
ST(m)			2	2	2	2	2	2
MTS(q)							2	2
MATR		✓			✓	✓		✓
DSC				✓		✓		✓
				Misspecification				
$\Theta^\star = \emptyset$ in sample	No	No	No	No	No	No	No	No
p-value for $H_0 : \Theta^\star \neq \emptyset$								
				Bounds				
$\mathrm{SD}^+_{\mathrm{avg}}$	.000	.000	.025	.025	.027	.027	.025	.027
	.947	.947	.933	.933	.933	.933	.380	.380
$\mathrm{SD}^+_{\mathrm{avg}}(\cdot\|0)$	.000	.000	.193	.193	.193	.193	.193	.193
	.581	.581	.576	.576	.576	.576	.560	.560
$\mathrm{SD}^+_{\mathrm{avg}}(\cdot\|00)$	.000	.000	.335	.335	.335	.335	.335	.335
	1.00	1.00	.992	.992	.992	.992	.965	.965
$\mathrm{SD}^+_{\mathrm{avg}}(\cdot\|1)$	.000	.000	.010	.010	.010	.010	.010	.010
	.970	.970	.966	.966	.966	.966	.371	.371
$\mathrm{SD}^+_{\mathrm{avg}}(\cdot\|11)$	.000	.000	.010	.010	.010	.010	.010	.010
	1.00	1.00	.996	.996	.996	.996	.383	.383
$\mathrm{SD}_{\mathrm{avg}}$	.000	.000	.054	.054	.054	.054	.054	.054
	1.00	1.00	.976	.976	.976	.976	.423	.423

References

Abbring, J. H. and J. J. Heckman (2007): "Chapter 72 Econometric Evaluation of Social Programs, Part III: Distributional Treatment Effects, Dynamic Treatment Effects, Dynamic Discrete Choice, and General Equilibrium Policy Evaluation," in *Handbook of Econometrics*, ed. by J. J. Heckman and E. E. Leamer, Elsevier, vol. Volume 6, Part 2, 5145–5303.

Alessie, R., S. Hochguertel, and A. v. Soest (2004): "Ownership of Stocks and Mutual Funds: A Panel Data Analysis," *The Review of Economics and Statistics*, 86, 783–796.

Andrews, D. W. K. and P. J. Barwick (2012): "Inference for Parameters Defined by Moment Inequalities: A Recommended Moment Selection Procedure," *Econometrica*, 80, 2805–2826.

Andrews, D. W. K. and S. Han (2009): "Invalidity of the bootstrap and the m out of n bootstrap for confidence interval endpoints defined by moment inequalities," *Econometrics Journal*, 12, S172–S199.

Andrews, D. W. K. and G. Soares (2010): "Inference for Parameters Defined by Moment Inequalities Using Generalized Moment Selection," *Econometrica*, 78, 119–157.

Angrist, J. D. and G. M. Kuersteiner (2011): "Causal Effects of Monetary Shocks: Semiparametric Conditional Independence Tests with a Multinomial Propensity Score," *Review of Economics and Statistics*, 93, 725–747.

Arulampalam, W., A. Booth, and M. Taylor (2000): "Unemployment persistence," *Oxford Economic Papers*, 52, 24–50.

Balke, A. and J. Pearl (1994): "Counterfactual Probabilities: Computational Methods, Bounds, and Applications," in *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI-94)*, ed. by R. Lopez de Mantras and D. Poole, 46–54.

——— (1997): "Bounds on Treatment Effects From Studies With Imperfect Compliance," *Journal of the American Statistical Association*, 92, 1171–1176.

Bartolucci, F. and V. Nigro (2010): "A Dynamic Model for Binary Panel Data With Unobserved Heterogeneity Admitting a n-Consistent Conditional Estimator," *Econometrica*, 78, 719–733.

Bernard, A. B. and J. B. Jensen (2004): "Why Some Firms Export," *The Review of Economics and Statistics*, 86, 561–569.

Bhuller, M., C. N. Brinch, and S. Konigs (2016): "Time Aggregation and State Dependence in Welfare Receipt," *Econ J*, n/a–n/a.

Bonhomme, S. (2012): "Functional Differencing," *Econometrica*, 80, 1337–1385.

Browning, M. and J. M. Carro (2010): "Heterogeneity in dynamic discrete choice models," *Econometrics Journal*, 13, 1–39.

——— (2014): "Dynamic binary outcome models with maximal heterogeneity," *Journal of Econometrics*, 178, 805–823.

Bugni, F. A. (2010): "Bootstrap Inference in Partially Identified Models Defined by Moment Inequalities: Coverage of the Identified Set," *Econometrica*, 78, 735–753.

Bugni, F. A., I. A. Canay, and X. Shi (2017): "Inference for subvectors and other functions of partially identified parameters in moment inequality models," *Quantitative Economics*, 8, 1–38.

Canay, I. A. (2010): "EL inference for partially identified models: Large deviations optimality and bootstrap validity," *Journal of Econometrics*, 156, 408–425.

Canay, I. A. and A. M. Shaikh (2017): "Practical and Theoretical Advances in Inference for Partially Identified Models," in *Advances in Economics and Econometrics*, ed. by B. Honore, A. Pakes, M. Piazzesi, and L. Samuelson, Cambridge University Press, 271–306.

Card, D. and R. Hyslop (2005): "Estimating the Effects of a Time-Limited Earnings Subsidy for Welfare-Leavers," *Econometrica*, 73, 1723–1770.

Carro, J. M. (2007): "Estimating dynamic panel data discrete choice models with fixed effects," *Journal of Econometrics*, 140, 503–528.

Chamberlain, G. (1984): "Chapter 22 Panel data," in *Handbook of Econometrics*, ed. by Z. Griliches and M. D. Intriligator, Elsevier, vol. Volume 2, 1247–1318.

——— (1985): "Heterogeneity, OMitted Variable Bias, and Duration Dependence," in *Longitudinal Analysis of Labor Market Data*, ed. by J. Heckman and B. Singer, Cambridge University Press.

——— (2010): "Binary Response Models for Panel Data: Identification and Information," *Econometrica*, 78, 159–168.

Chay, K. Y., H. Hoynes, and D. Hyslop (2004): "True State Dependence in Monthly Welfare Participation: A Nonexperimental Analysis," *Working paper*.

Chen, X., E. Tamer, and A. Torgovitsky (2011): "Sensitivity Analysis in Semiparametric Likelihood Models," *Cowles Foundation Discussion Paper 1836*.

Chernozhukov, V., I. Fernández-Val, J. Hahn, and W. Newey (2013): "Average and Quantile Effects in Nonseparable Panel Models," *Econometrica*, 81, 535–580.

Chernozhukov, V., H. Hong, and E. Tamer (2007): "Estimation and Confidence Regions for Parameter Sets in Econometric Models," *Econometrica*, 75, 1243–1284.

Chernozhukov, V., W. Newey, and A. Santos (2015): "Constrained conditional moment restriction models," *cemmap working paper CWP 59/15*.

Chetty, R. (2008): "Moral Hazard versus Liquidity and Optimal Unemployment Insurance," *Journal of Political Economy*, 116, 173–234.

Chiburis, R. C. (2010): "Semiparametric bounds on treatment effects," *Journal of Econometrics*, 159, 267–275.

Christensen, B. J., R. Lentz, D. T. Mortensen, G. R. Neumann, and A. Werwatz (2005): "On-the-Job Search and the Wage Distribution," *Journal of Labor Economics*, 23, 31–58.

Connault, B. (2016): "Hidden Rust Models," *Working paper*.

Contoyannis, P., A. M. Jones, and N. Rice (2004): "The Dynamics of Health in the British Household Panel Survey," *J. Appl. Econ.*, 19, 473–503.

Corcoran, M. and M. S. Hill (1985): "Reoccurrence of Unemployment among Adult Men," *The Journal of Human Resources*, 20, 165–183.

Demuynck, T. (2015): "Bounding average treatment effects: A linear programming approach," *Economics Letters*, 137, 75–77.

Deza, M. (2015): "Is there a stepping stone effect in drug use? Separating state dependence from unobserved heterogeneity within and between illicit drugs," *Journal of Econometrics*, 184, 193–207.

Domencich, T. and D. L. McFadden (1975): *Urban Travel Demand: A Behavioral Analysis*, North-Holland Publishing Co.

Drakos, K. and P. T. Konstantinou (2013): "Investment decisions in manufacturing: assessing the effects of real oil prices and their uncertainty," *J. Appl. Econ.*, 28, 151–165.

Dubé, J.-P., G. J. Hitsch, and P. E. Rossi (2010): "State dependence and alternative explanations for consumer inertia," *The RAND Journal of Economics*, 41, 417–445.

Ellwood, D. T. (1982): "Teenage unemployment: Permanent scars or temporary blemishes?" in *The youth labor market problem: Its nature, causes, and consequences*, University of Chicago Press, 349–390.

Embrechts, P. and M. Hofert (2013): "A note on generalized inverses," *Mathematical Methods in Operations Research*, 77, 423–432–.

Eriksson, S. and D.-O. Rooth (2014): "Do Employers Use Unemployment as a Sorting Criterion When Hiring? Evidence from a Field Experiment," *American Economic Review*, 104, 1014–39.

Faberman, R. J., A. Mueller, A. Şahin, and G. Topa (2017): "Job Search Behavior among the Employed and Non-Employed," Tech. rep.

Fang, Z., T. Hu, and H. Joe (1994): "On the Decrease in Dependence with Lag for Stationary Markov Chains," *Probability in the Engineering and Informational Sciences*, 8, 385–401.

Farber, H., C. Herbst, D. Silverman, and T. von Wachter (2018): "Whom Do Employers Want? The Role of Recent Employment and Unemployment Status and Age," Tech. rep.

Farber, H. S., J. Rothstein, and R. G. Valletta (2015): "The Effect of Extended Unemployment Insurance Benefits: Evidence from the 20122013 Phase-Out," *American Economic Review*, 105, 171176.

Farber, H. S., D. Silverman, and T. von Wachter (2016): "Determinants of Callbacks to Job Applications: An Audit Study," *American Economic Review*, 106, 314–318.

Farber, H. S., D. Silverman, and T. M. von Wachter (2017): "Factors Determining Callbacks to Job Applications by the Unemployed: An Audit Study," *RSF: The Russell Sage Foundation Journal of the Social Sciences*, 3, 168.

Farber, H. S. and R. G. Valletta (2015): "Do Extended Unemployment Benefits Lengthen Unemployment Spells?: Evidence from Recent Cycles in the U.S. Labor Market," *Journal of Human Resources*, 50, 873–909.

Fernández-Val, I. (2009): "Fixed effects estimation of structural parameters and marginal effects in panel probit models," *Journal of Econometrics*, 150, 71–85.

Flinn, C. and J. Heckman (1982): "New methods for analyzing structural models of labor force dynamics," *Journal of Econometrics*, 18, 115–168.

Fourer, R., D. M. Gay, and B. W. Kernighan (2002): *AMPL: A Modeling Language for Mathematical Programming*, Cengage Learning.

Freyberger, J. and J. L. Horowitz (2015): "Identification and shape restrictions in non-parametric instrumental variables estimation," *Journal of Econometrics*, 189, 41–53.

Ghayad, R. (2013): "The Jobless Trap," *Working paper*.

Grogger, J. (2004): "Welfare transitions in the 1990s: The economy, welfare policy, and the EITC," *J. Pol. Anal. Manage.*, 23, 671–695.

Gurobi Optimization, I. (2015): "Gurobi Optimizer Reference Manual," .

Ham, J. C., D. Iorio, and M. Sovinsky (2013): "Caught in the Bulimic Trap?: Persistence and State Dependence of Bulimia Among Young Women," *Journal of Human Resources*, 48, 736–767.

Ham, J. C. and L. Shore-Sheppard (2005): "The effect of Medicaid expansions for low-income children on Medicaid participation and private insurance coverage: evidence from the SIPP," *Journal of Public Economics*, 89, 57–83.

Handel, B. R. (2013): "Adverse Selection and Inertia in Health Insurance Markets: When Nudging Hurts," *American Economic Review*, 103, 2643–82.

Hansen, L. P., J. Heaton, and E. G. J. Luttmer (1995): "Econometric Evaluation of Asset Pricing Models," *The Review of Financial Studies*, 8, 237–274.

Heckman, J. J. (1978): "Simple Statistical Models for Discrete Panel Data Developed and Applied to Test the Hypothesis of True State Dependence against the Hypothesis of Spurious State Dependence," *Annales de l'inséé*, 227–269.

——— (1981a): "Heterogeneity and State Dependence," in *Studies in Labor Markets*, ed. by S. Rosen, University of Chicago Press.

——— (1981b): "Statistical Models for Discrete Panel Data," in *Structural Analysis of Discrete Data and Econometric Applications*, ed. by C. F. Manski and D. McFadden, MIT Press: Cambridge, MA.

——— (1981c): "The Incidental Parameters Problem and the Problem of Initial Conditions in Estimating a Discrete Time-Discrete Data Stochastic Process," in *Structural Analysis of Discrete Data with Econometric Applications*, ed. by C. F. Manski and D. McFadden, MIT Press: Cambridge, MA.

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

Heckman, J. J. and S. Navarro (2007): "Dynamic discrete choice and dynamic treatment effects," *Journal of Econometrics*, 136, 341–396.

Heckman, J. J., J. Smith, and N. Clements (1997): "Making the Most Out of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Programme Impacts," *The Review of Economic Studies*, 64, 487–535.

Heckman, J. J. and E. J. Vytlacil (2007): "Chapter 70 Econometric Evaluation of Social Programs, Part I: Causal Models, Structural Models and Econometric Policy Evaluation," in *Handbook of Econometrics*, ed. by J. J. Heckman and E. E. Leamer, Elsevier, vol. Volume 6, Part 2, 4779–4874.

Heckman, J. J. and R. J. Willis (1977): "A Beta-logistic Model for the Analysis of Sequential Labor Force Participation by Married Women," *Journal of Political Economy*, 85, 27–58.

Honoré, B. (2002): "Nonlinear models with Panel Data," *Portuguese Economic Journal*, 1, 163–179.

Honoré, B. E. and E. Kyriazidou (2000): "Panel Data Discrete Choice Models with Lagged Dependent Variables," *Econometrica*, 68, 839–874.

Honoré, B. E. and A. Lewbel (2002): "Semiparametric Binary Choice Panel Data Models without Strictly Exogeneous Regressors," *Econometrica*, 70, 2053–2063.

Honoré, B. E. and E. Tamer (2006): "Bounds on Parameters in Panel Dynamic Discrete Choice Models," *Econometrica*, 74, 611–629.

Hopenhayn, H. A. and E. C. Prescott (1992): "Stochastic Monotonicity and Stationary Distributions for Dynamic Economies," *Econometrica*, 60, 1387–1406.

Hu, Y. and M. Shum (2012): "Nonparametric identification of dynamic models with unobserved state variables," *Journal of Econometrics*, 171, 32–44.

Hyslop, D. R. (1999): "State Dependence, Serial Correlation and Heterogeneity in Intertemporal Labor Force Participation of Married Women," *Econometrica*, 67, 1255–1294.

IBM (2010): *IBM ILOG AMPL Version 12.2*, International Business Machines Corporation.

Irace, M. (2018): "Patient Loyalty in Hospital Choice: Evidence from New York," *Working paper*.

Jarosch, G. and L. Pilossoph (forthcoming): "Statistical Discrimination and Duration Dependence in the Job Finding Rate," *The Review of Economic Studies*.

Judge, G. G., W. E. Griffiths, R. C. Hill, H. Lütkepohl, and T.-C. Lee (1985): *The Theory and Practice of Econometrics*, Wiley, second ed.

Kaido, H., F. Molinari, and J. Stoye (2016): "Inference on Projections of Identified Sets," *Working paper*.

Kasahara, H. and K. Shimotsu (2009): "Nonparametric Identification of Finite Mixture Models of Dynamic Discrete Choices," *Econometrica*, 77, 135–175.

Keane, M. P. (1997): "Modeling Heterogeneity and State Dependence in Consumer Choice Behavior," *Journal of Business & Economic Statistics*, 15, 310–327.

Keane, M. P., P. E. Todd, and K. I. Wolpin (2011): "Chapter 4 - The Structural Estimation of Behavioral Models: Discrete Choice Dynamic Programming Methods and Applications," Elsevier, vol. Volume 4, Part A, 331–461.

Kitamura, Y. and J. Stoye (2013): "Nonparametric Analysis of Random Utility Models: Testing," *cemmap working paper CWP36/13*.

Kroft, K., F. Lange, and M. J. Notowidigdo (2013): "Duration Dependence and Labor Market Conditions: Evidence from a Field Experiment," *The Quarterly Journal of Economics*, 128, 1123–1167.

Lafférs, L. (2013): "Essays in Partial Identification," Ph.D. thesis, Norwegian School of Economics.

——— (2018): "Bounding average treatment effects using linear programming," *Empirical Economics*.

Lehmann, E. L. (1966): "Some Concepts of Dependence," *The Annals of Mathematical Statistics*, 37, 1137–1153.

Magnac, T. (2000): "Subsidised Training and Youth Employment: Distinguishing Unobserved Heterogeneity from State Dependence in Labour Market Histories," *The Economic Journal*, 110, 805–837.

Magnac, T. and D. Thesmar (2002): "Identifying Dynamic Discrete Decision Processes," *Econometrica*, 70, 801–816.

Manski, C. (1994): "The selection problem," in *Advances in Econometrics, Sixth World Congress*, vol. 1, 143–70.

Manski, C. F. (1997): "Monotone Treatment Response," *Econometrica*, 65, 1311–1334.

——— (2006): "Two Problems of Partial Identification with Panel Data," 13th International Conference on Panel Data, July 7–9, Cambridge.

——— (2007): "Partial Indentification of Counterfactual Choice Probabilities," *International Economic Review*, 48, 1393–1410.

——— (2014): "Identification of income-leisure preferences and evaluation of income tax policy," *Quantitative Economics*, 5, 145–174.

Manski, C. F. and J. V. Pepper (2000): "Monotone Instrumental Variables: With an Application to the Returns to Schooling," *Econometrica*, 68, 997–1010.

——— (2009): "More on monotone instrumental variables," *Econometrics Journal*, 12, S200–S216.

Mogstad, M., A. Santos, and A. Torgovitsky (2018a): "Using Instrumental Variables for Inference about Policy Relevant Treatment Parameters," *Econometrica (forthcoming)*.

Mogstad, M., A. Torgovitsky, and C. R. Walters (2018b): "Identification of Causal Effects with Multiple Instruments: Problems and Some Solutions," *Working paper*.

Molinari, F. (2008): "Partial identification of probability distributions with misclassified data," *Journal of Econometrics*, 144, 81–117.

Mourifié, I. (2015): "Sharp bounds on treatment effects in a binary triangular system," *Journal of Econometrics*, 187, 74–81.

Mühleisen, M. and K. F. Zimmermann (1994): "A panel analysis of job changes and unemployment," *European Economic Review*, 38, 793–801.

Narendranathan, W. and P. Elias (1993): "Influences of past history on the incidence of youth unemployment: empirical findings for the UK," *Oxford Bulletin of Economics and Statistics*, 55, 161–185.

Norets, A. (2009): "Inference in Dynamic Discrete Choice Models With Serially orrelated Unobserved State Variables," *Econometrica*, 77, 1665–1682.

Norets, A. and X. Tang (2014): "Semiparametric Inference in Dynamic Binary Choice Models," *The Review of Economic Studies*, 81, 1229–1262.

Nunley, J. M., A. Pugh, N. Romero, and R. A. Seals (2016): "The Effects of Unemployment and Underemployment on Employment Opportunities," *ILR Review*, 70, 642–669.

Oberholzer-Gee, F. (2008): "Nonemployment stigma as rational herding: A field experiment," *Journal of Economic Behavior & Organization*, 65, 30–40.

Okumura, T. and E. Usui (2014): "Concave-monotone treatment response and monotone treatment selection: With an application to the returns to schooling," *Quantitative Economics*, 5, 175–194.

Pakes, A. and J. Porter (2016): "Moment Inequalities for Multinomial Choice with Fixed Effects," Tech. rep.

Prowse, V. (2012): "Modeling Employment Dynamics With State Dependence and Unobserved Heterogeneity," *Journal of Business & Economic Statistics*, 30, 411–431.

Romano, J. P. and A. M. Shaikh (2008): "Inference for identifiable parameters in partially identified econometric models," *Journal of Statistical Planning and Inference*, 138, 2786–2807.

——— (2010): "Inference for the Identified Set in Partially Identified Econometric Models," *Econometrica*, 78, 169–211.

Rudin, W. (1976): *Principles of mathematical analysis*, New York: McGraw-Hill.

Rust, J. (1994): "Chapter 51 Structural estimation of markov decision processes," in *Handbook of Econometrics*, ed. by R. F. Engle and D. L. McFadden, Elsevier, vol. Volume 4, 3081–3143.

Shaikh, A. M. and E. J. Vytlacil (2011): "Partial Identification in Triangular Systems of Equations With Binary Dependent Variables," *Econometrica*, 79, 949–955.

Shaked, M. and J. G. Shanthikumar (2007): *Stochastic orders*, Springer.

Shapiro, A. and D. Dentcheva (2014): *Lectures on stochastic programming: modeling and theory*, vol. 16, SIAM.

Stokey, N. L., R. E. Lucas, and E. C. Prescott (1989): *Recursive Methods in Economic Dynamics*, Harvard University Press.

Torgovitsky, A. (forthcoming): "Partial Identification by Extending Subdistributions," *Quantitative Economics*.

Tumino, A. (2015): "The scarring effect of unemployment from the early '90s to the Great Recession," *Working paper*.

Wooldridge, J. M. (2005): "Simple solutions to the initial conditions problem in dynamic, nonlinear panel data models with unobserved heterogeneity," *J. Appl. Econ.*, 20, 39–54.

——— (2010): *Econometric analysis of cross section and panel data*, MIT press.