# Partial Identification of State Dependence

Alexander Torgovitsky*

February 12, 2015

**Abstract**

This paper is about the empirical measurement of state dependence in dynamic binary outcomes. Most of the literature on this topic focuses on the estimation of parametric dynamic binary response (DBR) panel data models. Identification in these models requires extensive assumptions about functional form, heterogeneity and the exogeneity of covariates. In contrast, I focus on what can be learned from the data under easily interpretable nonparametric assumptions. To do this, I propose a dynamic potential outcomes (DPO) model and develop nonparametric counterparts of the parameters and assumptions considered in the traditional DBR literature. I show how to construct sharp identified sets in the DPO model using a flexible linear programming procedure that is valid for a large variety of parameters and auxiliary identifying assumptions. Confidence regions for these identified sets are obtained by applying recent results from the literature on inference in moment inequality models. The analysis is applied to study state dependence in the labor force participation of married women. Using conservative, nonparametric assumptions, it is possible to reject the hypothesis that there is no state dependence in the labor force participation outcomes of married women.

**JEL classification:** C14; C20; C51; J2; J6

**Keywords:** State dependence, partial identification, linear programming, moment inequalities, intertemporal labor force participation

*Department of Economics, Northwestern University, a-torgovitsky@northwestern.edu. This paper was presented at the University of Wisconsin at Madison. I thank the audience at that seminar. My thanks also to Ivan Canay for several helpful discussions, and to Joachim Freyberger, Yuichi Kitamura, Chuck Manski, Matt Notowidigdo and Jack Porter for providing useful feedback.

## 1 Introduction

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

As observed by Heckman and Willis (1977), Heckman (1978, 1981a) and many subsequent authors, positive serial correlation in the observed employment outcomes $Y_i \equiv (Y_{i0}, Y_{i1}, \ldots, Y_{iT})$ (conditional on observed covariates) is not necessarily an indication of state dependence. An alternative explanation is that agents have some persistent latent heterogeneity in their propensities for employment and, as a result, some agents are always more likely to be employed or non-employed than other agents. This mechanism 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 long-run efficacy of active labor market programs designed to increase employment (Heckman, 1978, 1981a). It is therefore important to have convincing econometric methods to quantify the role of state dependence in the observed persistence in employment. In order to be convincing, these econometric methods must first address the identification problem of distinguishing state dependence from persistent unobserved heterogeneity. This paper proposes and analyzes a new framework for thinking about this identification problem with observational data.[1]

Perhaps the most widely adopted empirical approach for measuring state dependence with observational data involves the estimation of some variant of a dynamic binary response (DBR) model. As I discuss more in Section 2, the identification of these threshold-crossing style models depends on a large number of unpalatable as-

[1]For experimental evidence on state (and duration) dependence in employment outcomes see the recent studies by Oberholzer-Gee (2008), Kroft et al. (2013), Ghayad (2013) and Eriksson and Rooth (2014). Quasi-experimental evidence on state dependence is more rare, but see Lee (2008), who proposes a regression discontinuity design to determine the causal effect of political incumbency on elections outcomes, and Handel (2013), who uses the institutional structure of health insurance choice, together with employee turnover, to identify state dependence (or "inertia") in these choices.

sumptions about the data generating process, including arbitrary shape restrictions on the distribution of heterogeneity. These assumptions are motivated by analytic convenience rather than economic theory, which makes their credibility suspect. As a result, estimates from parametric DBR models may not be convincing measures of state dependence.

The main contribution of this paper is the development of a new nonparametric framework for identifying state dependence in observational data. Instead of attempting to modify a threshold-crossing model, I propose a dynamic potential outcomes (DPO) model that describes the causal effect of previous outcomes on current outcomes. The model is simple to state: Given a binary outcome $Y_{it} \in \{0, 1\}$ for agent $i$ at time $t$, it posits the existence of two latent variables $U_{it}(0)$ and $U_{it}(1)$ that represent the outcome 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 (or structure) is the joint distribution across all time periods of $U_{it}(0)$ and $U_{it}(1)$, together with the initial period observed outcome $Y_{i0}$. From knowledge of this structure, one can construct a number of interesting measures of state dependence, including (but not limited to) commonly used measures such as the average treatment effect.[2]

Static potential outcomes models have enjoyed widespread adoption by economists for cross-sectional applications. Their attraction lies in their fundamentally nonparametric description of the causal process, which promotes empirical analysis based on transparent and easily interpretable assumptions. This benefit is shared in the dynamic extension proposed in this paper, which is entirely nonparametric at its core. An additional benefit in the dynamic setting is that the potential outcomes framework allows for general patterns of observed and unobserved heterogeneity, while also permitting complex temporal dependence structures among the latent factors that affect outcomes.

Measures of state dependence in the DPO model are in general not point identified. I derive sharp worst-case bounds that use only the empirical evidence. These bounds are very wide and show that empirical evidence alone cannot reject the possibility of no state dependence. Maintaining additional non-data (auxiliary) identifying assumptions leads to smaller identified sets. I propose and analyze several such assumptions. However, due to the dynamic nature of the model, it is typically difficult

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

to derive analytic expressions for identified sets under additional assumptions. Instead, I develop a general procedure for computing sharp identified sets of scalar parameters that incorporate the identifying content of these auxiliary identifying assumptions. In many cases, this computational method amounts to solving two linear programming problems and is therefore straightforward to implement. One attractive feature of this approach is its flexibility, which allows the analyst to choose parameters and impose assumptions based on their economic rationale, rather than their mathematical expediency. I propose several types of parameters that may be of interest depending on the empirical question and the analyst's goals or methodological preferences.

Confidence regions that account for sampling variation are obtained by recasting the DPO model as a moment inequality model. This allows for the application of recent results from the literature about statistical inference in moment inequality models, see e.g. Andrews and Soares (2010) and the references cited therein. Most of this literature has focused on constructing confidence regions for the entire parameter vector through test inversion. In the DPO model, only scalar or low-dimensional functions of the entire parameter vector are of ultimate interest. In principle, confidence regions for the entire parameter vector can be projected down to confidence regions for low-dimensional parameters. However, doing so requires first constructing the confidence region for the entire parameter vector. In the DPO model, the entire parameter vector typically has dimension in the thousands or ten-thousands, so constructing its confidence region is computationally infeasible. To address this problem, I apply the recent results of Bugni et al. (2014) that show how one can construct uniformly valid confidence regions for scalar or low-dimensional parameters in a computationally straightforward manner by effectively profiling the GMS procedure. I also consider a profiled subsampling approach proposed by Romano and Shaikh (2008) and a "minimum quantile" statistic that combines both methods. A Monte Carlo study provides some evidence on the finite-sample efficacy of these procedures as applied to the DPO model.

The econometric framework proposed in this paper can be applied to any of the large variety of empirical problems in which identifying state dependence is important. These include the dynamics of welfare recipiency (Chay et al., 2004; Card and Hyslop, 2005), product choices among consumers (Keane, 1997; Dubé et al., 2010; Handel, 2013), self-reported health status (Contoyannis et al., 2004), firm investment (Drakos and Konstantinou, 2013) and exporting (Bernard and Jensen, 2004) decisions, household investment behavior (Alessie et al., 2004), illicit drug usage (Deza, 2015), and eating disorders (Ham et al., 2013).

To focus the analysis, I apply the methodology to the previously discussed problem of measuring state dependence in the labor force participation of married women.

Following previous authors (Hyslop, 1999; Keane and Sauer, 2009), I use a balanced panel drawn from the Panel Study of Income Dynamics (1986) (PSID) that consists of 1,812 women observed yearly between 1979 and 1985. I use the application to illustrate the identifying content of various auxiliary identifying assumptions and discuss their justification. I show that the hypothesis of no state dependence can be rejected by maintaining weak, nonparametric assumptions that impose stationarity and monotonicity. Under stronger (but still nonparametric) assumptions about stationarity and the sign of dynamic selection, I estimate that 4.1–45.3% of married women are affected by state dependence in the sense that they would be employed in a given year if and only if they were employed in the previous year. This confidence interval is consistent with (but wider than) that obtained from a standard parametric DBR model.

The organization of the paper is as follows. In the next section, I provide a brief review of a type of DBR model commonly used in the literature, as well as a brief review of the recent research on nonparametric models of dynamic discrete outcomes. In Section 3, I develop and analyze the dynamic extension of the potential outcomes model as an alternative approach to measuring state dependence. This section is an abstract description of the methodology. In Section 4, the methodology is discussed more concretely in the context of measuring state dependence in the labor force participation of married women in the PSID data. I describe methods for conducting statistical inference in Section 5, report the results of a Monte Carlo study, and present confidence intervals for the PSID data. Section 6 contains some concluding remarks.

## 2 Dynamic Binary Response Models

A commonly used econometric tool for detecting state dependence for binary outcomes in the presence of heterogeneity is the parametric DBR model.[3] A textbook version of the model (e.g. Wooldridge (2010)) specifies the threshold-crossing equation

$$Y_{it} = \mathbb{1}[\gamma Y_{i(t-1)} + X_{it}'\beta + \lambda Y_{i0} + A_i + V_{it} \geq 0] \quad \text{for } t \geq 1, \tag{1}$$

[3]Many of the empirical papers listed in the introduction use this model or a closely related variant of it. Linear probability models are also occasionally used to analyze state dependence in discrete outcomes (e.g. pp. 1265–1266 of Hyslop (1999)), however they have highly undesirable properties when viewed as models of heterogeneous treatment response, see Manski and Pepper (2009) pg. S210, so I do not consider them here.

The parametric DBR model discussed here is myopic in the sense that it is not based on the optimizing behavior of a forward-looking agent, in contrast to the sorts of models discussed by (e.g.) Rust (1994). Structural models of this sort tend to have difficulty accommodating persistent latent heterogeneity except in very rudimentary ways.

where both $A_i$ and $V_{it}$ are unobservables and $X_{it}$ is a vector of observed covariates. As in many other panel data models, the unobservables are divided into a time-invariant component $A_i$ and a time-varying component $V_{it}$. Including the initial period outcome $Y_{i0}$ as an explanatory variable was proposed by Wooldridge (2005) as a simple solution to the initial conditions problem observed by Heckman (1981b). A baseline set of assumptions placed on (1) includes the following:

***A1.*** $V_{it}|X_i, Y_{i0}, A_i \sim N(0,1)$ *for all* $t$*, where* $X_i \equiv (X_{i0}, X_{i1}, \ldots, X_{iT})$*, and* $\text{Cov}(V_{it}, V_{is}) = 0$ *for all* $s \neq t$*.*

***A2.*** $A_i|X_i, Y_{i0} \sim N(0, \sigma_A^2)$*.*

Assumptions A1 and A2 together with (1) enable the construction of a likelihood function for $(Y_{i1}, \ldots, Y_{iT})$, conditional on $X_i, Y_{i0}$. If A1, A2 and (1) are valid, then the corresponding maximum likelihood estimator of $(\gamma, \beta, \lambda, \sigma_A^2)$ will be consistent and asymptotically normal under reasonable regularity conditions. From consistent estimators of these parameters, one can construct a consistent estimator of the average treatment effect of $Y_{i(t-1)}$ on $Y_{it}$, i.e.

$$\text{ATE} \equiv \mathbb{E}\left[\mathbb{1}[\gamma + X_{it}'\beta + \lambda Y_{i0} + A_i + V_{it} \geq 0] - \mathbb{1}[X_{it}'\beta + \lambda Y_{i0} + A_i + V_{it} \geq 0]\right], \quad (2)$$

see (e.g.) Wooldridge (2005) for a clear exposition. Interest in the ATE defined in (2) implicitly presupposes that (1) is a causal model capable of describing the value that $Y_{it}$ would obtain under exogenous manipulations in $Y_{i(t-1)}$. The ATE is a reasonable measure of state dependence given the constraints of the DBR model. As I discuss more in Section 4, it is not obvious that it is the parameter one would be interested in when working with a more flexible model.

If A1, A2 or (1) are incorrect, the maximum likelihood estimator based on these assumptions will generally be inconsistent for the estimated parameters, leading also to inconsistency in the resulting estimator of the ATE. For analyses concerned with fit or prediction, this may not be an important issue. However, when attempting to ascribe a causal interpretation to a parameter like the ATE, it is of paramount importance that the model is not badly misspecified. That a number of compelling criticisms of A1, A2 and (1) have been raised in the literature should therefore be cause for considerable caution.

One frequently discussed criticism is the treatment of $A_i$ as a random effect in A2. In linear panel data models, time-invariant unobservables like $A_i$ can be treated as fixed effects and removed through differencing transformations. The distribution of $A_i$ does not need to be parametrically specified, and its dependence with the ob-

served explanatory variables can be left unrestricted. In nonlinear specifications such as (1), differencing does not eliminate $A_i$. Moreover, if $T$ is small then treating the $A_i$ as parameters to be estimated through maximum likelihood creates an incidental parameters problem that pollutes estimates of the other parameters.

Several ways of addressing this problem have been discussed in the literature. Chamberlain (1984) proposed replacing A2 with the correlated random effects assumption that $A_i|X_i, Y_{i0} \sim N(\mu_0 Y_{i0} + \sum_{t=1}^{T} X'_{it}\mu_t, \sigma_A^2)$, so that the mean of $A_i$ can vary with $(X_i, Y_{i0})$, thereby allowing for some limited dependence between $A_i$ and $(X_i, Y_{i0})$. Rasch (1961) and Andersen (1970) discovered that if the normal distribution for $V_{it}$ in A1 is replaced by a logistic distribution, then there exists a non-linear transformation of the outcome probabilities that eliminates the $A_i$. This result was extended considerably by Honoré and Kyriazidou (2000); see also Bonhomme (2012) for a unifying analysis. Honoré and Lewbel (2002) showed that in the presence of a special regressor $A_i$ can be treated as a fixed effect and several other parts of Assumptions A1 and A2 can be relaxed. However, their results only identify the parameter coefficients and not the ATE. Fernández-Val (2009) has revisited the incidental parameters problem created by treating each $A_i$ as a true fixed effect to be estimated. He argues that the bias induced on the ATE by estimating the incidental parameters may be relatively small even for small $T$, and proposes a bias-corrected estimator. Carro (2007) shows that the bias stemming from the incidental parameters problem can be mitigated by employing a modified maximum likelihood estimator.

These solutions to the incidental parameters problem still maintain substantial parametric assumptions about the distributions on $V_{it}$ or $A_i$ (or both). Arguably, the logistic solution even amplifies the importance of correctly specifying the distribution of $V_{it}$, since the validity of this approach relies crucially on the logistic function form (Chamberlain, 2010). Such parameterizations are used because they enable the construction of a finitely-parameterized likelihood function, which addresses the question of identification, at least as a mathematical problem. Many researchers are skeptical of models identified only through the force of arbitrary parameterizations, see e.g. Manski (1975) for an early criticism.

This paper is premised on the view that this skepticism is justified. Economic theory rarely suggests functional forms for the distributions of latent variables. In parametric DBR models, these functional forms are chosen for analytical and computational convenience, rather than compelling economic rationale. The sensitivity of empirical conclusions to these assumptions is difficult to characterize rigorously. For these reasons, economists conducting empirical research increasingly favor identification arguments based on intuitive, nonparametric assumptions about the treatment

assignment and outcome determination processes.

Some researchers have responded to this view by investigating identification in nonparametric dynamic models of binary outcomes. Browning and Carro (2010, 2014) assume that $Y_{it}$ follows a homogenous, first-order Markov process, conditional on $A_i$, i.e. that $Y_{it}|Y_{i(t-1)} = \overline{y}, \ldots, Y_{i0}, A_i$ is distributed like $Y_{i1}|Y_{i0} = \overline{y}, A_i$ for all $t$.[4] They show that a nonparametric counterpart to the ATE can be point identified (sometimes only locally) if $A_i$ is assumed to be discretely distributed with a sufficiently small support of known dimension. Kasahara and Shimotsu (2009) derive different sufficient conditions for point identification in closely related models, and consider the additional identifying power of excluded exogenous covariates. Their analysis also requires $A_i$ to be discretely distributed with support of known size. Hu and Shum (2012) and Shiu and Hu (2013) maintain similar assumptions while allowing $A_i$ to be continuously distributed. Their conditions for identification include high-level completeness assumptions which can be difficult to interpret and/or verify in applications. In addition, the first-order conditional Markov property assumed by all of these papers may be unattractive for employment outcomes, see Section 6 of Browning and Carro (2014).

These papers show that for dynamic binary outcomes it is difficult to achieve nonparametric point identification under easily interpretable conditions while still allowing for general forms of unobserved heterogeneity. Given this difficulty, it seems sensible to entertain a partial identification approach of the sort advocated by Manski (2003). To the best of my knowledge, the only other authors to consider partial identification for models of dynamic binary outcomes are Honoré and Tamer (2006), Chernozhukov et al. (2013), Pakes and Porter (2014) and Norets and Tang (2014).

The latter paper considers partial identification of a semiparametric structural model of dynamic binary decision making. These models have the benefit of being directly derived from a theoretical model of rational forward-looking decision making. However, their complexity necessarily requires some strong and undesirable assumptions even in the relatively agnostic framework considered by Norets and Tang (2014).[5] Structural models such as these are useful tools for considering the impacts of counterfactual policy interventions with no historical precedent. The goal of measuring causal effects that is pursued in this paper is more modest in comparison, but the methods employed will maintain fewer, and more easily interpretable assumptions. The analysis

[4] Additional covariates $X_{it}$ are included in this nonparametric framework by simply conditioning, so I suppress them in the notation when natural.

[5] In particular, many analyses of structural dynamic binary choice models maintain an assumption of no persistent unobserved heterogeneity. The results of the aforementioned papers by Kasahara and Shimotsu (2009) and Hu and Shum (2012) can be used to provide identification conditions for structural models with some forms of persistent heterogeneity, potentially under strong additional conditions.

in this paper should therefore be viewed as complementary to Norets and Tang (2014) and the literature on structural models of dynamic binary outcomes.[6]

More related to this paper is the work of Honoré and Tamer (2006), who showed how to construct identified sets for parametric DBR models that maintain A1 but relax A2 to allow $A_i$ to be a finitely-distributed fixed effect.[7] Chernozhukov et al. (2013) extended these results to allow for $A_i$ to be continuously distributed asymptotically. Another important contribution of Chernozhukov et al. (2013) was to derive non-sharp bounds on the ATE for the nonparametric model

$$Y_{it} = g(Y_{i(t-1)}, A_i, V_{it}) \tag{3}$$

under the assumption that $V_{it}|Y_{i(t-1)}, \ldots, Y_{i0}, A_i$ is distributed identically to $V_{i1}|Y_{i0}, A_i$ for all $t$. Equation (3) can be seen as a nonparametric, nonseparable counterpart to (1), while the assumption on $V_{it}$ can be seen as a counterpart to the stationarity and serial independence assumptions embedded in A1. Chernozhukov et al. (2013) describe this assumption as "time is an instrument."[8]

This paper presents an alternative approach to partial identification of state dependence for binary outcomes. It is motivated by the observation that (3) is an unnecessarily rich model for measuring state dependence. Instead, one can limit attention to the latent random variables

$$U_{it}(0) = g(0, A_i, V_{it}) \quad \text{and} \quad U_{it}(1) \equiv g(1, A_i, V_{it}). \tag{4}$$

From the distribution of $U_i \equiv (Y_0, U_{i1}(0), \ldots, U_{iT}(0), U_{i1}(1), \ldots, U_{iT}(1))$ one can construct common parameters of interest, such as the ATE at time $t$, i.e. $\mathbb{E}[U_{it}(1) - U_{it}(0)] = \mathbb{E}[g(1, A_i, V_{it}) - g(0, A_i, V_{it})]$. The distribution of $U_i$ is also sufficient to determine the implied distribution of observed outcomes, since by construction $Y_{it} = U_{it}(Y_{i(t-1)}) = g(Y_{i(t-1)}, A_i, V_{it})$ for $t \geq 1$. Hence, a model of $U_i$ is both complete—in the sense of generating a distribution of the observed endogenous variables $Y_i$—and sufficiently rich to answer common causal questions about state dependence.

[6] See also Heckman and Navarro (2007), who consider point identification in dynamic structural models using identification-at-infinity style arguments.

[7] Honoré and Tamer (2006) also did not condition on $Y_{i0}$, thereby allowing for a partial identification treatment of the initial conditions problem.

[8] Recently, Pakes and Porter (2014) have shown how a condition similar to the "time is an instrument" assumption can be combined with a separable index structure for $g$ to construct non-sharp identified sets for parameters in this index structure. Their analysis is semiparametric in that they do not impose finite-dimensional parameterizations for the distributions of latent variables. However, their results do not suggest bounds on causal parameters (such as the ATE), which are the focus of this paper.

The benefit of considering a model based on (4) instead of the more complicated model in (3) is that it is much easier to characterize identified sets in (4). As I demonstrate in the next section, this is a consequence of the discreteness of $U_i$, which enables the construction of identified sets for a variety of parameters under a variety of assumptions through linear programming. Indeed, for the model of (3) with the "time is an instrument" assumption, Chernozhukov et al. (2013) only provided indirect arguments to suggest that their proposed bounds were sharp. Formally showing sharpness in their model is quite difficult due to the large number of possible joint distributions of $A_i$ and $V_{it}$. This complication is removed by limiting attention to $U_i$, which is a collection of binary random variables. A practical consequence is that an analyst working with (4) is afforded drastically more freedom in selecting parameters of interest and auxiliary identifying assumptions, while still having a method for computing sharp identified sets.

The DPO model is formally defined in the next section. Instead of motivating it as a simpler version of (3), I describe it as a dynamic extension of the static potential outcomes models. Both descriptions are useful; the former as a comparison with the important work of Chernozhukov et al. (2013), and the latter because it allows me to use insights from the enormous literature on static potential outcomes models.

## 3 The Dynamic Potential Outcomes Model

### 3.1 Model and Definitions

The canonical static potential outcomes model postulates the existence of 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 observed treatment state $D_i$ and the potential outcomes through $Y_i = D_i U_i(1) + (1 - D_i)U_i(0)$.[9] The goal of the analysis is to recover an informative feature (mean, quantile, etc.) of the distribution $U_i(1) - U_i(0)$ of treatment effects from the observable distribution of $(Y_i, D_i)$.

This paper is concerned with understanding the causal effect of lagged outcomes on current and future outcomes. At each time $t = 1, \ldots, T$, the outcome is $Y_{it}$ and the "treatment" is the immediately preceding outcome, $Y_{i(t-1)}$. I assume throughout the main text that $Y_{it} \in \{0, 1\}$ is binary for each $t$, so that at time $t$ both $Y_{it}$ and $Y_{i(t-1)}$ are binary.[10] Hence, in analogy to the static potential outcomes model, suppose that

[9] Standard practice is to denote $U_i(0)$ and $U_i(1)$ by $Y_i(0)$ and $Y_i(1)$. This turns out to be somewhat confusing in the dynamic model, which is why I use $U_i(0), U_i(1)$ to denote potential outcomes.

[10] The analysis is extended to multi-valued outcomes in Appendix A.

for each time period $t = 1, \ldots, T$ there exist 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 in time $t$ had the past period outcome $Y_{i(t-1)}$ been exogenously manipulated to be 0 or 1, respectively. A causal interpretation of the parametric DBR model of the previous section implicitly defines (suppressing covariates)

$$U_{it}(y) = \mathbb{1}[\gamma y + \lambda Y_{i0} + A_i + V_{it} \geq 0]. \tag{5}$$

The DPO model does not impose this type of linear index structure on $U_{it}(0)$ and $U_{it}(1)$.

The observed outcomes $Y_i \equiv (Y_{i0}, Y_{i1}, \ldots, Y_{iT})$ together form a random $(T+1)$–vector with values in $\mathcal{Y} \equiv \{0, 1\}^{T+1}$. The observed outcomes are related to the vectors of 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

$$Y_{it} = Y_{i(t-1)}U_{it}(1) + (1 - Y_{i(t-1)})U_{it}(0) = U_{it}(Y_{i(t-1)}) \quad \text{for all } t \geq 1. \tag{6}$$

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 (1981b) by simply reducing the number of observed variables that are explicitly modeled, similar in spirit to the approach of Wooldridge (2005) for parametric DBR models.[11]

Embedded in this specification is the presumption that the analyst is only interested in the causal effect of the outcome in period $t-1$ on the outcome in period $t$. In some settings, it may be interesting to analyze the causal effects of specific sequences of previous outcomes on the current period outcome. This can be accommodated by redefining the collection of 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.

In addition to $Y_i$, the analyst 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, however I assume throughout the analysis that $\mathcal{X}$ is a finite set, i.e. $X$ is discretely distributed.[12] Some of the components of $X_{it}$ may be thought of as

[11] In particular, note that the DPO model *does not* assume that $Y_{i0}$ is independent of any of the potential outcomes. In principle, this could be added later as an auxiliary identifying assumption, but this is rarely justifiable (Heckman, 1981b).

[12] The partial identification analysis discussed in the next section extends to continuously distributed $X$ in a straightforward way. Using the terminology introduced in that section, the easiest way to accommodate continuous $X$ is to define the structures as conditional-on-$[X = x]$ probability mass functions for $U$. However, this conditional-on-$[X = x]$ formulation presents additional challenges for statistical inference, so for

conditioning variables that serve to capture observed heterogeneity, while others might be considered as instruments that satisfy certain exclusion or monotonicity conditions in relationship to the potential outcomes. Examples of these types of assumptions are discussed in detail in Section 4.3.

The DPO model captures state dependence in a natural way through the possibility that $U_{it}(0) \neq U_{it}(1)$. That is, the outcome $Y_{it} = U_{it}(Y_{i(t-1)})$ for agent $i$ that actually occurred in period $t$ may have been different had $Y_{i(t-1)}$ been different. The model allows for observed and unobserved heterogeneity quite generally. Observed heterogeneity is reflected through differences in the distributions of $(U_i(0), U_i(1))|X_i = x$ for different values of $x$. These conditional–on–$[X_i = x]$ distributions need not be degenerate, which allows for additional unobserved heterogeneity. 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 baseline model discussed in this section does not separate this unobserved heterogeneity into permanent and transitory components like the models in Section 2. As a consequence, no restrictions are imposed on the dynamic correlation of the potential outcomes. While it is possible to include a distinction between permanent and transitory components by imposing the bounds implied by the "time is an instrument" assumption of Chernozhukov et al. (2013) (see Section 4), this is by no means necessary or essential to the DPO model.

### 3.2 Partial Identification

This section describes a general procedure for constructing identified sets in the DPO model. The analysis is presented abstractly here in the sense that it is only assumed that there is a particular parameter of interest and that the data generating process satisfies a certain set of auxiliary identifying assumptions. The next two sections contain concrete examples of parameters of interest and auxiliary identifying assumptions tailored to the problem of distinguishing state dependence in female labor force participation. I assume throughout the analysis that the available panel is balanced with periods indexed by $t = 0, 1, \ldots, T$ for $T$ small and fixed, and that it is i.i.d. across agents $i = 1, \ldots, n$. For notational simplicity, I drop the $i$ subscript until discussing statistical inference in Section 5.

Using language similar to Hurwicz (1950), a structure for the dynamic potential outcomes model maintaining (6) 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

consistency in exposition I do not explicitly consider it here.

$U \equiv (Y_0, U(0), U(1))$.[13] 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

$$\sum_{u \in \mathcal{U}, x \in \mathcal{X}} P[u,x] = 1. \tag{7}$$

Let $\mathcal{P}$ denote the set of all functions $P : \mathcal{U} \times \mathcal{X} \to [0,1]$ that satisfy (7).

The admissible set is the subset $\mathcal{P}^\dagger$ of $\mathcal{P}$ to which the analyst restricts attention. In practice, $\mathcal{P}^\dagger$ will be composed of the structure $P$ that satisfy auxiliary identifying assumptions, examples of which are discussed in Section 4.3. As a formalization, it is convenient to assume that $\mathcal{P}^\dagger = \{P \in \mathcal{P} : \rho(P) \geq 0\}$, where $\rho : \mathcal{P} \to \mathbb{R}^{d_\rho}$ is a function representing restrictions on $P$, and the inequality is interpreted componentwise. Equality restrictions can be incorporated into $\mathcal{P}^\dagger$ by including pairs of inequalities in the function $\rho$. The restrictions may also depend on features of the observable distribution of $(Y, X)$, but this is suppressed in the notation until Section 5 when the distinction becomes salient.

The identified set, denoted by $\mathcal{P}^\star$, is defined as the subset of the admissible set $\mathcal{P}^\dagger$ that could have generated the observed data through relationship (6). Let $\mathbb{P}[Y = \cdot, X = \cdot]$ denote the observable probability mass function of $(Y, X)$, where $Y \equiv (Y_0, Y_1, \ldots, Y_T)$. Then $P \in \mathcal{P}^\star$ requires that for every $y \equiv (y_0, y_1, \ldots, y_T) \in \mathcal{Y}$ and $x \in \mathcal{X}$,

$$\mathbb{P}[Y = y, X = x] = \mathbb{P}_P[Y = y, X = x] = \mathbb{P}_P[Y_0 = y_0, U_t(y_{t-1}) = y_t \text{ all } t \geq 1, X = x],$$

where $\mathbb{P}_P[\cdot]$ denotes the probability of an event when $(U, X)$ is distributed according to $P$. This expression can be rewritten as a linear function of $\{P[u,x] : u \in \mathcal{U}, x \in \mathcal{X}\}$:

$$\mathbb{P}[Y = y, X = x] = \sum_{u \in \mathcal{U}_{\text{oeq}}(y)} P[u,x], \tag{8}$$

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 (8) for $T = 2$.

Observe that (8) 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 (7) and $1 \geq P[u,x] \geq 0$ for all $u \in \mathcal{U}$, $x \in \mathcal{X}$. Hence, if $\rho$ is also a linear function of $P$, then determining whether a given $P$ is in the identified set is equivalent to determining the existence of a solution to a system of linear equations. This is a well-studied problem

[13] Throughout the paper, the notation $\{0,1\}^s$ refers to the $s$–fold Cartesian product of the set $\{0,1\}$.

Figure 1: Observational Equivalence, $T = 2$

| Potential Outcomes | | | | | Observed Outcomes | | |
|---|---|---|---|---|---|---|---|
| $Y_0$ | $U_1(0)$ | $U_1(1)$ | $U_2(0)$ | $U_2(1)$ | $Y_0$ | $Y_1$ | $Y_2$ |
| $\boxed{0}$ | $\boxed{0}$ | 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 |
| | $\vdots$ | | $\vdots$ | | | $\vdots$ | |

The full diagram would have $2^{2T+1} \equiv 2^5 = 32$ rows corresponding to all possible realizations of potential outcomes. Here, the rows shown are those corresponding to those potential outcomes that could generate $Y = (0, 0, 0)$ or $Y = (1, 0, 1)$ through the recursive relationship (6), i.e. the elements of $\mathcal{U}_{\text{oeq}}(0, 0, 0)$ and $\mathcal{U}_{\text{oeq}}(1, 0, 1)$ in (8). The observed realization of $Y$ provides knowledge of the potential outcomes that are boxed, but not those that are not boxed.

for which fast and reliable computational solutions exist. In Section 4.3, I provide several examples of identifying assumptions that can be represented as functions $\rho$ that are linear in $P$.

A probability mass function $P$ is typically too complex of an object to be of ultimate interest. Instead, an analyst is usually interested in the identified set $\Theta^\star \equiv \{\theta(P) : P \in \mathcal{P}^\star\}$ for a lower-dimensional feature (parameter) $\theta : \mathcal{P} \to \mathbb{R}^{d_\theta}$ of $P$. A standard example of a parameter $\theta$ is the average treatment effect at time $t$, $\text{ATE}_t(P) \equiv \mathbb{E}_P[U_t(1) - U_t(0)]$, where $\mathbb{E}_P$ denotes expectation taken with respect to $P$. Although it is suppressed in the notation, $\theta$ can depend on features of the distribution of observables $(Y, X)$.

In general, $\Theta^\star$ can be traced out by determining for any candidate $t$ in the range of $\theta$ whether there exists a function $P \in \mathcal{P}^\star$ such that $\Theta(P) = t$. If $\theta$ is scalar-valued, then the identified set can, in many situations, be determined by solving two optimization problems.

**Theorem 1.** *Suppose that $\mathcal{P}^\dagger$ is closed and convex, and that $\theta$ is a continuous, scalar-*

*valued function of $P$. Then, as long as $\mathcal{P}^\star$ is nonempty, $\Theta^\star = [\theta_l^\star, \theta_u^\star]$, where*

$$\theta_l^\star \equiv \min_{P \in \mathcal{P}^\star} \theta(P) = \min_{\{P[u,x] \in [0,1] : u \in \mathcal{U}, x \in \mathcal{X}\}} \theta(P) \text{ s.t. } \rho(P) \geq 0, \text{ (7), and (8) } \forall y, x$$

$$\text{and } \theta_u^\star \equiv \max_{P \in \mathcal{P}^\star} \theta(P) = \max_{\{P[u,x] \in [0,1] : u \in \mathcal{U}, x \in \mathcal{X}\}} \theta(P) \text{ s.t. } \rho(P) \geq 0, \text{ (7), and (8) } \forall y, x.$$

***Proof of Theorem 1.*** 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 (8) for all $y$ and $x$. The image of the continuous, real-valued function $\theta$ over this closed, convex (and non-empty) set is a closed non-empty interval (e.g. Theorem 4.22 of Rudin (1976)) with smallest value $\theta_l^\star$ and largest value $\theta_u^\star$, i.e. $\Theta^\star = [\theta_l^\star, \theta_u^\star]$. *Q.E.D.*

If $\rho$ and $\theta$ are linear, then determining whether a given $t$ is in $\Theta^\star$ is equivalent to determining the existence of a solution to a system of linear equations. If $\theta$ is also scalar, then the two optimization problems in Theorem 1 are linear programs. In Section 4.2, I discuss several examples of interesting choices of $\theta$ that are both scalar and linear in $P$.

The traditional approach to building identified sets is to propose bounds for a given parameter and then construct an admissible structure under which the extreme points of these bounds are obtained. This strategy was employed in the pioneering work of Manski (1989) as well as more recent work by Chesher (2010), Shaikh and Vytlacil (2011) and Khan et al. (2011), among many others. It has the benefit of providing analytic expressions for the bounds, which can often yield useful intuition as to the source and strength of identification. Analytic expressions can also aid the construction of valid confidence regions.

However, in more complicated models, it quickly becomes difficult or impossible to explicitly construct sharp identified sets, especially when the admissible set is defined by many restrictions. In this case, the broad applicability of results like Theorem 1 are attractive. In particular, Theorem 1 easily allows one to change parameters and identifying assumptions without requiring a lengthy and potentially difficult re-derivation of the analytical formulas that characterize the identified set. This general point about partial identification analysis has been appreciated (sometimes implicitly) by many other authors, including Honoré and Tamer (2006), Manski (2007), Molinari (2008), Chiburis (2010), Kitamura and Stoye (2013), Freyberger and Horowitz (2013), Manski (2014) and Lafférs (2015). The latter work uses a similar computational strategy as in this paper for a static potential outcomes model. However, the benefits in that 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.

### 3.3 Computation and Dimension Reduction

The optimization problem in Theorem 1 can be quite large. For example, the application to female labor force participation in Section 4 has $T = 6$ so that even without covariates the dimension of $P = \{P[u, x] : u \in \mathcal{U}, x \in \mathcal{X}\}$ is $2^{2T+1} = 8{,}192$. A nonparametric specification employed in Section 4 takes $\mathcal{X}$ to be a set with 20 support points, increasing the overall number of variables in the problem to $20 \times 8{,}192 = 163{,}840$. The number of constraints in the problem—even without any auxiliary identifying assumptions—is at least $2^{T+1} \times |\mathcal{X}| = 128 \times 20 = 2{,}560$ for (8), plus $2 \times 163{,}840$ constraints to ensure that $P$ is contained in the unit interval.

These dimensions are large for a general optimization problem. However, if both $\rho$ and $\theta$ are linear so that the optimization problem is a linear program, then these dimensions are actually quite modest. A standard desktop computer with sophisticated linear programming algorithms can finish the above problem with a few hundred thousand variables and constraints in well under a minute.[14] Since many of the constraints can be expected to be redundant, it is important to use software with algebraic capabilities such as AMPL (Fourer et al., 2002) to achieve this type of speed. Programming languages like AMPL also make it straightforward to change parameters (objective functions) and add or remove auxiliary identifying assumptions (constraints), which allows an analyst to easily exploit the flexibility of Theorem 1.

Still, in situations when $T$ is large or $X$ assumes many values, the dimensions of the linear programs in Theorem 1 may be prohibitive. A semiparametric specification can be used to address the dimensionality problems caused by $X$ assuming many values—this is currently under development. For situations where $T$ and hence $\mathcal{U}$ are large, one solution is to limit the analysis to agents that had less than a certain number of transitions in the time horizon under consideration, where a transition is defined as occurring when $Y_t \neq Y_{t-1}$. In many data sets, such as the female labor force participation data analyzed in the following section, the overwhelming majority of the observed units have less than 2 or 3 transitions over the time horizon in the data—see Table 1. By removing the small subset of the population that has more than (say) 3 transitions, one effectively restricts any observationally equivalent $P = \{P[u, x] : u \in \mathcal{U}, x \in \mathcal{X}\}$ to be 0 for any $u$ that generates an observed sequence with 4 or more transitions. This can characterize a large proportion of potential outcomes $u \in \mathcal{U}$, even though only a small proportion of the population has more than 4 transitions.[15]

[14] All optimization problems in this paper were solved using KNITRO 9 (Byrd et al., 2006).

[15] Strictly speaking, this affects the interpretation of the parameter of interest in the same way that selecting a sample affects the interpretation of the parameter of interest. One can either be content with the new interpretation or formally characterize the effect of out-of-sample extrapolation using the worst-case

By requiring these $P[u, x]$ to be 0, they are essentially removed from the optimization problem, thereby reducing the number of remaining variables.

Note that the computational concerns with large $T$ are not unique to the DPO model. Rather, it seems to be a characteristic of 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 assumption that $V_t$ is serially uncorrelated in A1 is replaced by the assumption that $V_t$ follows an AR(1) process. As observed by Heckman (1981a) and Chamberlain (1984), this implies that $Y_t|A$ is not Markov of any order. The resulting likelihood function for the parametric DBR involves a $T$–dimensional integral, which becomes increasingly difficult to evaluate (or simulate) as $T$ grows.

## 4 State Dependence in Female Labor Force Participation

### 4.1 Background and Data

Economists have long been interested in explaining the determinants and dynamics of female labor force participation. Early work by Gronau (1974) and Heckman (1974) emphasized the selection problem inherent in analyzing a static cross-section of wages for working women. Heckman and Willis (1977) considered a dynamic model of female labor force participation that allowed for unobserved heterogeneity but not state dependence. Heckman (1978, 1981a,b) extended these analyses to allow for both unobserved heterogeneity and state dependence, and emphasized the identification difficulties inherent in such models. Structural approaches have been employed by Heckman and MaCurdy (1980), Eckstein and Wolpin (1989) and Eckstein and Lifshitz (2011), among others. The aim of the current section is to complement these works by providing estimates of state dependence under easily interpretable nonparametric assumptions, while also illustrating the DPO model discussed in the previous section.

I revisit the topic of female labor force participation using a dataset originally constructed by Hyslop (1999) and re-analyzed subsequently by Keane and Sauer (2009). Briefly, the sample is taken from the 1986 Panel Study of Income Dynamics (PSID) and consists of $n = 1{,}812$ women who were aged 18 – 60 in 1980 and continuously married to an employed husband between 1979 and 1985. Data is observed yearly between 1979 and 1985, so the initial period ($t = 0$) corresponds to 1979 and the terminal period ($t = T = 6$) corresponds to 1985. Following Hyslop (1999), the outcome variable $Y_t$ is specified as 1 if a woman reports both positive hours worked and positive earnings in year $t$. Hence, women with $Y_t = 0$ are either not participating in the labor force,

bounds approach of Manski (1996, Section IV).

or are unemployed participants. I refer to $Y_t = 0$ as non-employment. The analysis is concerned with state dependence in employment relative to non-employment, but does not distinguish between non-participation and unemployment.[16]

The available observed covariates ($X$) are the same as in Hyslop (1999): permanent nonlabor income, transitory nonlabor income in each period, number of children aged 0-2, 3-5 and 6-17 in each period, age, highest reported level of education over the sample period, and race (black/non-black). Permanent nonlabor income is defined as the average of the husband's log earnings over the sample period. Transitory nonlabor income in each period is defined as the deviation of the husband's log earnings in that period from permanent nonlabor income. The time-varying covariates are transitory nonlabor income and number of children. See Hyslop (1999) and Keane and Sauer (2009) for a complete discussion of the covariates. Table 1 provides some descriptive statistics concerning the outcome process $Y_t$.

As a baseline point of comparison for the discussion ahead, column (P) of Table 2S reports the point estimate of the ATE from a parametric DBR model. In particular, .24 in column (P) is the point estimate of the ATE as constructed from the maximum likelihood estimator for model (1) under A1 and A2. Following the specification in Hyslop (1999), the components of $X_t$ contain age, age squared, race, highest reported level of education over the sample period, permanent income (as defined above), transitory income in each time period and number of children of ages 0-2, 3-5 and 6-17 in each period. In addition, the same correlated random effects specification as in Hyslop (1999) is used to allow for some limited dependence between $X$ and $A$—see that paper for details. A 95% bootstrapped confidence interval for the ATE is $[.144, .337]$.

### 4.2 Parameters of Interest

A natural measure of state dependence is the proportion of women that would have experienced a different employment outcome in period $t$ had their employment outcome at $t-1$ been different, i.e. the proportion of women with the event $[U_t(0) \neq U_t(1)]$. In the binary outcome cases considered here, such affected agents are characterized for any given period $t$ by $[U_t(0) = 0, U_t(1) = 1]$ or $[U_t(0) = 1, U_t(1) = 0]$. The proportion of the first group under structure $P$ is denoted by

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

[16] To stay consistent with the literature I will still refer to this problem as one of female labor force *participation* despite the misnomer.

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. These are the women who would have been employed in period $t$ had they (exogenously) been employed in period $t-1$, but would be non-employed in period $t$ had they been non-employed in period $t-1$. The measure of the second group under structure $P$ is denoted by

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

This is the proportion of women who would have experienced negative state dependence in employment at time $t$ in the sense that they would be non-employed in period $t$ if and only if they had been employed in period $t-1$. The total proportion of women experiencing state dependence under structure $P$ is $\mathrm{SD}_t^+(P) + \mathrm{SD}_t^-(P)$.

The observed data do not directly point identify $\mathrm{SD}_t^+$ or $\mathrm{SD}_t^-$ for the same usual reasons as in the static potential outcomes models. First, an analyst never observes both $U_t(0)$ and $U_t(1)$, since only $Y_t = U_t(Y_{t-1})$ is observed. Hence, quantities like $\mathrm{SD}_t^+$ that concern the joint distribution of $(U_t(0), U_t(1))$ are inherently not point identified. Second, even the marginal distributions of $U_t(0)$ and $U_t(1)$ will generally not be point identified due to the endogeneity of prior outcomes. That is, in general we expect that for observationally equivalent $P$,

$$\mathbb{P}[Y_t = 1 | Y_{t-1} = 1, X] = \mathbb{P}_P[U_t(1) = 1 | Y_{t-1} = 1, X] \neq \mathbb{P}_P[U_t(1) = 1 | X], \tag{9}$$

since $Y_{t-1} = 1$ depends on $(U_{t-1}(0), U_{t-1}(1))$, and $U_{t-1}(1)$ is likely correlated with $U_t(1)$, even conditional on $X$, due to permanent unobserved heterogeneity.

Instead, there are a range of values of $\mathrm{SD}_t^+$ and $\mathrm{SD}_t^-$ that are compatible with the data, i.e. these parameters are partially identified. To determine the identified sets for these parameters, first note that they are linear functions of $P = \{P[u, x] : u \in \mathcal{U}, x \in \mathcal{X}\}$ since

$$\mathrm{SD}_t^+(P) = \sum_{u \in \mathcal{U}_t^+} \left( \sum_{x \in \mathcal{X}} P[u, x] \right), \tag{10}$$

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 $u_t(1) = 1$. A similar linear expression can be derived for $\mathrm{SD}_t^-$. This linearity means that Theorem 1 can be applied to quickly and reliably compute the identified sets for these parameters

through linear programming if $\rho$ is also linear.[17] These computed identified sets (for $t = 3$, i.e. 1982) are given in column (1) of Table 2S. Lengths of all identified sets are shown in Table 2L. Since no additional assumptions are being placed on the potential outcomes (i.e. $\mathcal{P}^\dagger = \mathcal{P}$), these bounds can be described as using only the empirical evidence.

The empirical-evidence-only bounds are large. The width of the identified set for $\text{SD}_t^+$ is .87—not much smaller than the a priori largest width of 1. In Appendix C, it is shown that the sharp empirical-evidence-only identified set for $\text{SD}_t^+$ is given by

$$[0,\ \mathbb{P}[Y_{t-1} = 0, Y_t = 0] + \mathbb{P}[Y_{t-1} = 1, Y_t = 1]]\,. \tag{11}$$

The intuition behind these bounds is that agents with $[Y_{t-1} = 0, Y_t = 1]$ or $[Y_{t-1} = 1, Y_t = 0]$ cannot be those with potential outcomes $[U_t(0) = 0, U_t(1) = 1]$, since in the first case $U_t(0) = 1$, while in the second $U_t(1) = 0$. On the other hand, agents with $[Y_{t-1} = 0, Y_t = 0]$ could have potential outcomes given by either $[U_t(0) = 0, U_t(1) = 0]$ or $[U_t(0) = 0, U_t(1) = 1]$. It is the second of these groups that experiences positive state dependence, and their proportion of the overall population could be 0, but can be no greater than $\mathbb{P}[Y_t = 0, Y_{t-1} = 0]$. Similarly, the observable group characterized by $[Y_{t-1} = 1, Y_t = 1]$ is comprised of agents with $[U_t(0) = 0, U_t(1) = 1]$ or $[U_t(0) = 1, U_t(1) = 1]$. The first of these groups experiences positive state dependence and their proportion of the overall population could be 0, but it can be no greater than $\mathbb{P}[Y_{t-1} = 1, Y_t = 1]$. Combined, the proportion of the population with positive state dependence at time $t$ can be no greater than the upper bound in (11). This bound is large when observed outcomes have strong positive serial correlation, which is the case for the data studied here—see Table 1.

Using analogous reasoning, it can be shown (Appendix C) that sharp bounds on $\text{SD}_t^-$ are given by

$$[0,\ \mathbb{P}[Y_{t-1} = 0, Y_t = 1] + \mathbb{P}[Y_{t-1} = 1, Y_t = 0]]\,. \tag{12}$$

The empirical-evidence-only upper bounds on negative state dependence are large when the observed outcomes have strong negative serial correlation. Intuitively, more than a small amount of negative state dependence would imply frequent transitions in labor force participation, while the actual PSID data contains relatively few such transitions. Hence, using only the empirical evidence, it is possible to rule out the hypothesis that more than a small number of women (13%) experience a negative causal effect of past

[17] Here $\rho$ is currently null since no additional assumptions have been imposed, i.e. $\mathcal{P}^\dagger = \mathcal{P}$.

employment on future employment.

Since the lower bounds in (11) and (12) are sharp and always 0 regardless of the distribution of the data, the empirical evidence alone never enables a rejection of no state dependence. The non-existence of state dependence can only be established by incorporating auxiliary identifying assumptions, such as those discussed ahead in Section 4.3.

In the parametric DBR model of Section 2, state dependence is governed by the scalar parameter $\gamma$. It is often argued (e.g. Wooldridge (2010)) that analysts interested in state dependence should focus on the average treatment effect given in (2) that results from changing $Y_{t-1}$ from 0 to 1 and viewing (1) as a causal model. Recent nonparametric studies such as Kasahara and Shimotsu (2009), Chernozhukov et al. (2013) and Browning and Carro (2014) have followed this lead of considering the ATE as the parameter of primary interest.

In the DPO framework it is still possible to consider the average treatment effect $\text{ATE}_t(P) \equiv \mathbb{E}_P[U_t(1)] - \mathbb{E}_P[U_t(0)]$ as a measure of state dependence. However, it is also possible to consider $\text{SD}_t^+$ and $\text{SD}_t^-$, which may be more interesting to an analyst, depending on their goals. Unlike $\text{ATE}_t$, $\text{SD}_t^+$ depends on the joint distribution of $(U_t(0), U_t(1))$ and not just the marginals. It therefore captures the treatment effect within the population, rather than the difference in treatment distributions between two subpopulations. This constitutes an example of the distinction between the difference of potential outcome distributions and the distribution of potential outcome differences; see e.g. Manski (1996, 1997b) and Heckman et al. (1997) for a discussion in a traditional static framework.

To see this point more clearly, observe that the relationship between $\text{SD}_t^+$ and $\text{ATE}_t$ is given by

$$\begin{aligned}\text{ATE}_t(P) &= (\mathbb{P}_P[U_t(1) = 1, U_t(0) = 0] + \mathbb{P}_P[U_t(1) = 1, U_t(0) = 1]) \\ &\quad - (\mathbb{P}_P[U_t(1) = 0, U_t(0) = 1] + \mathbb{P}_P[U_t(1) = 1, U_t(0) = 1]) \\ &= \text{SD}_t^+(P) - \text{SD}_t^-(P).\end{aligned}$$

Hence, $\text{ATE}_t$ is the proportion of the population that experiences positive state dependence, less the proportion that experiences negative state dependence. As a result, $\text{ATE}_t = \text{SD}_t^+$ if and only if $\text{SD}_t^- = 0$. In general, it is possible for $\text{ATE}_t$ to be small or zero even if there is both positive and negative state dependence.

Depending on the application, an analyst may be interested in both $\text{SD}_t^+$ and $\text{SD}_t^-$. For example, suppose that $Y_t$ denotes welfare status as in Chay et al. (2004) or Card and Hyslop (2005). Then $\text{SD}_t^-$ represents the 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 $\text{SD}_t^+$ represents the proportion of the population that are in the "welfare trap." On the other hand, for female labor force participation it may be reasonable to assume that $\text{SD}_t^-(P) = 0$ for all $P \in \mathcal{P}^\dagger$, in which case $\text{ATE}_t = \text{SD}_t^+$ by force of the assumption. In the next section I refer to this condition as monotone treatment response (following the terminology of Manski (1997a)), and I consider its effect on the identified set for $\text{SD}_t^+$ both alone and when combined with other assumptions. However, for different outcome variables, such as welfare recipiency, a monotonicity assumption like this may be unpalatable.

The distinction between $\text{SD}_t^+$, $\text{SD}_t^-$ and $\text{ATE}_t$ as parameters describing state dependence highlights one of the chief benefits of the empirical framework proposed in this paper. The discreteness of the DPO model enables the identified set to be computed easily for a variety of different parameters. This computational ease, combined with an acknowledgment that point identification can only be obtained under unpalatably strong assumptions, provides the freedom to choose parameters based on their relevance to the application, rather than their analytic tractability. This type of approach to empirical research is in harmony with what Heckman and Urzua (2010) have called "Marchak's Maxim."

The identified sets for parameters analogous to $\text{SD}_t^+$, $\text{SD}_t^-$ and $\text{ATE}_t$ can be computed for subgroups defined by realizations of $X$ and/or $Y$. For example, one might be interested in comparing the identified regions of

$$\text{SD}_t^+(P|x) \equiv \mathbb{P}_P[U_t(0) = 0, U_t(1) = 1|X = x] = \frac{\sum_{u \in \mathcal{U}_t^+} P[u, x]}{\mathbb{P}[X = x]}$$

for different values of $x \in \mathcal{X}$. This can provide a description of how state dependence varies across subgroups defined by different combinations of observables. These types of parameters are also easily studied using parametric or nonparametric binary response models, so the DPO model possess no additional advantage in this regard.

However, parameters that condition on components of $Y$ cannot be easily analyzed using other types of binary response models. For example, in studying female labor force participation, an analyst may be interested in identifying positive state dependence among just women who are currently non-employed. This parameter is given by

$$\text{SD}_t^+(P|0) \equiv \mathbb{P}_P[U_t(0) = 0, U_t(1) = 1|Y_t = 0].$$

It is straightforward to show (Appendix D) that $\text{SD}_t^+(\cdot|0)$ is still a linear function of $P$.

Hence, Theorem 1 can be applied to quickly and reliably compute the sharp identified set for this parameter.

The empirical-evidence-only identified set for $\text{SD}_t^+(\cdot|0)$ for $t = 3$ (1982) is given in the second row of column (1) in Table 2S. It turns out to be slightly narrower than the identified set for the unconditional positive state dependence parameter $\text{SD}_t^+$. Deriving an analytical expression for the identified set of $\text{SD}_t^+(\cdot|0)$ seems difficult. Essentially, one needs to find bounds on $\mathbb{P}_P[U_t(0) = 0, U_t(1) = 1, Y_t = 0]$, but this is complicated by the fact that $Y_t$ depends on the entire history of counterfactual outcomes through (6). This emphasizes the utility of the DPO framework and Theorem 1 in freeing the analyst to identify parameters that are relevant for their analysis, regardless of whether they are analytically convenient.

Conditioning on past outcomes can be extended to analyze the subgroup of women who have been non-employed for the previous $m$ periods. Column (1) of Table 2S reports the empirical-evidence-only bounds for

$$\begin{aligned} \text{SD}_t^+(P|00) &\equiv \mathbb{P}_P[U_t(0) = 0, U_t(1) = 1|Y_t = 0, Y_{t-1} = 0] \\ \text{and } \ \text{SD}_t^+(P|000) &\equiv \mathbb{P}_P[U_t(0) = 0, U_t(1) = 1|Y_t = 0, Y_{t-1} = 0, Y_{t-2} = 0] \end{aligned}$$

with $t = 3$. These parameters are also linear functions of $P$ (Appendix D). The empirical-evidence-only identified sets for these parameters are as wide as is logically possible. The auxiliary identifying assumptions described in the next section will make these identified sets much smaller.

It is of course possible to replace the event $Y_t = 0$ with $Y_t = 1$, if the analyst is interested in the "treated" group. Conditioning on mixed sequences, such as $[Y_t = 1, Y_{t-1} = 0]$ is also straightforward, although of less obvious interest. Again, this flexibility is a great advantage of the DPO framework over the parametric and nonparametric binary response models in Section 2, in which attention is invariably restricted to the unconditional average treatment effect, regardless of whether this is the most relevant parameter for the application at hand.

### 4.3 Auxiliary Identifying Assumptions

The empirical-evidence-only identified sets for parameters measuring positive state dependence are very large. In this section I propose additional (auxiliary) identifying assumptions that can be placed on the potential outcomes in order to narrow these bounds. These assumptions are implemented by including restrictions in the $\rho$ function of Section 3.2 and then applying Theorem 1. Maintaining more assumptions leads to smaller identified sets, but less convincing inference, a trade-off described by Manski

(2003) as "The Law of Diminishing Credibility." In keeping with the spirit of Manski's law, I will present results using assumptions that become gradually stronger.

### 4.3.1 Monotone Treatment Response

In the previous section, it was observed that $\text{ATE}_t(P) \neq \text{SD}_t^+(P)$ unless $\text{SD}_t^-(P) = 0$. In some applications, it may make sense to assume that $\text{SD}_t^-(P) = 0$ for all $P \in \mathcal{P}^\dagger$. This seems reasonable with regards to female labor force participation, in which it specifies that no woman would be employed in period $t$ as a direct result of being non-employed in period $t-1$. Another way of stating the assumption is that $U_t(1) \geq U_t(0)$ with probability one under any admissible structure $P$. Hence, the condition can be viewed as the monotone treatment response (MTR) assumption of Manski (1997a) applied to the dynamic model.

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

Assumption MTR, along with the other assumptions discussed ahead, is shown in Appendix D to place a linear restriction on $P = \{P[u,x] : u \in \mathcal{U}, x \in \mathcal{X}\}$. As a result, the identified set for any of the parameters discussed in the previous section can be computed using Theorem 1 under any combination of the assumptions discussed in this section. Column (2) of Table 2S reports identified sets when MTR is imposed. As expected, $\text{SD}_t^-$ is point identified by force of the assumption. The identified set for $\text{SD}_t^+$ is unchanged relative to the empirical–evidence–only bounds. As it turns out, MTR has substantial identifying content for positive state dependence parameters when combined with some of the other assumptions discussed ahead, but has no content by itself, at least for the PSID data.

The parametric DBR model imposes either MTR or its opposite, depending on the sign of $\gamma$. This is because that model treats $\gamma$ as a fixed (deterministic) quantity, implying that with probability 1 either

$$U_t(1) = \mathbb{1}[\gamma + X_t'\beta + \lambda Y_0 + A + V_t \geq 0] \geq \mathbb{1}[X_t'\beta + \lambda Y_0 + A + V_t \geq 0] = U_t(0) \quad (13)$$

or the opposite. Hence, MTR does not represent a substantial assumption when compared to the parametric DBR model. This should not be viewed as a justification of the assumption. While MTR seems reasonable for labor force participation, it should not necessarily be taken for granted in other applications.

### 4.3.2 Stationarity

Some degree of time-invariance (stationarity) is a natural assumption in panel data settings. Assuming that the past is like the future allows the empirical evidence from different time periods to be combined. In the parametric DBR model of Section 2, all of the parameters are time-constant and all of the unobservables are stationary. In the DPO model, stationarity assumptions can be introduced by restricting the joint distribution of $(U_t(0), U_t(1))$ to be invariant across $t \geq 1$. A stronger version of this restriction could maintain the same condition on multiple time periods, e.g. that the distribution of $(U_{t-1}(0), U_t(0), U_{t-1}(1), U_t(1))$ does not vary across $t \geq 2$. More generally, consider the following assumption.

***Assumption ST:*** *Let $m \geq 0$ be a non-negative integer chosen by the analyst and define $U_t^m(0) \equiv (U_{t-m}(0), \ldots, U_t(0))$ and $U_t^m(1) \equiv (U_{t-m}(1), \ldots, U_t(1))$ for $t \geq m+1$. Then for any $P \in \mathcal{P}^\dagger$, every $u^m(0), u^m(1)$, and every $s, t \geq m+1$,*

$$\mathbb{P}_P[U_s^m(0) = u^m(0), U_s^m(1) = u^m(1)] = \mathbb{P}_P[U_t^m(0) = u^m(0), U_t^m(1) = u^m(1)].$$

As stated, ST is a restriction on the marginal distribution of $U$ (vs. $(U, X)$) for admissible structures $P$. It is possible to modify the condition so that it is conditional on $[X = x]$ for some or all $x \in \mathcal{X}$. Imposing the assumption conditional on all $x \in \mathcal{X}$ is stronger than the stated version of ST but may in many cases be just as reasonable. To avoid redundancy, I will state all future assumptions in terms of the unconditional distribution of potential outcomes whenever appropriate. However, it should be understood that this is simply to avoid tedium, and not necessarily because the conditions are more or less desirable in that form.

Identified sets maintaining only ST with $m = 0, 1, 2$ are shown in columns (3)–(5) of Table 2S.[18] While still quite wide, the identified sets with $m = 1, 2$ for the positive state dependence parameters are interesting, because the lower bound becomes larger than 0. This provides simple, nonparametric evidence against the hypothesis that correlation in outcomes is caused solely by persistent unobserved heterogeneity. Setting $m = 3$ makes the linear program infeasible, implying that the restriction constrained $\{P[u, x] : u \in \mathcal{U}, x \in \mathcal{X}\}$ so much that it was not able to satisfy the observational equivalence condition (8). Whether this infeasibility is due to sampling variation requires statistical considerations that will be addressed in Section 5.

Column (6) of Table 2S reports identified sets when MTR is added to ST. While MTR had no identifying power on its own (column (2)), it can be seen in columns

[18] Note that ST (with any $m \geq 0$) implies that $\text{ATE}_t$, $\text{SD}_t^+$ and $\text{SD}_t^-$ do not vary across $t$.

(5)–(6) to have substantial content when combined with ST. This finding again serves to underscore the importance of a general computational approach like Theorem 1. It shows that the identified sets for the DPO model are determined by the complicated interactions of many different equalities and inequalities, suggesting that an analytic approach would be highly infeasible.

It is important to notice that ST generally does not imply any stationarity in the distribution of observed outcomes $Y_t$. To see this, consider what would be required for stationarity of the observed outcomes in a simple stylized case in which $\mathbb{P}[Y_t = 0] = \mathbb{P}[Y_t = 1] = 1/2$ and $(U_{t+1}(0), U_{t+1}(1))$ is independent of $Y_t$. For any observationally equivalent $P$ satisfying these restrictions one has

$$\begin{aligned}
\mathbb{P}[Y_{t+1} = 1] &= \mathbb{P}[Y_{t+1} = 1, Y_t = 0] + \mathbb{P}[Y_{t+1} = 1, Y_t = 1] \\
&= \mathbb{P}[Y_t = 1] + \mathbb{P}[Y_{t+1} = 1, Y_t = 0] - \mathbb{P}[Y_{t+1} = 0, Y_t = 1] \\
&= \mathbb{P}[Y_t = 1] + \mathbb{P}_P[U_{t+1}(0) = 1, Y_t = 0] - \mathbb{P}_P[U_{t+1}(1) = 0, Y_t = 1] \\
&= \mathbb{P}[Y_t = 1] + \frac{1}{2}\Big( \mathbb{P}_P[U_{t+1}(0) = 1] - \mathbb{P}_P[U_{t+1}(1) = 0] \Big).
\end{aligned}$$

Hence, $\mathbb{P}[Y_{t+1} = 1] = \mathbb{P}[Y_t = 1]$ if and only if $\mathbb{P}_P[U_{t+1}(0) = 1] = \mathbb{P}_P[U_{t+1}(1) = 0]$. The latter condition does not restrict the distribution of potential outcomes across time. Rather, it is a statement about the joint distribution of potential outcomes at time $t + 1$. Even in its strongest forms, ST does not in any way impose such a restriction on the admissible $P$. Structures satisfying ST can therefore still generate observable distributions of $Y$ that are non-stationary.

#### 4.3.3 Diminishing Serial Correlation

Persistent heterogeneity in the propensity to be employed is likely to cause the potential outcomes to be positively serially correlated. However, it is reasonable 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:*** *Every $P \in \mathcal{P}^{\dagger}$ is such that for $y \in \{0, 1\}$, $\text{Corr}_P(U_t(y), U_{t+s}(y))$ is decreasing in $|s|$ for $s \in 1 - t, \ldots, T - t$.*

In general, DSC places a nonlinear restriction on $P$ and is therefore difficult to implement using Theorem 1. However, if ST holds (with any $m \geq 0$) then DSC becomes a linear restriction, equivalent to the statement that $\mathbb{P}_P[U_t(y) = 1, U_{t+s}(y) = 1]$ is

decreasing in $|s|$ for $s \in \{1-t, \ldots, T-t\}$ (see Appendix D for justification).[19] In light of this computational consideration, I will only consider the identifying content of DSC when it is combined with ST.

If $U_t(y)$ is determined through the threshold-crossing relationship (5), as in the parametric DBR model, and if ST holds, then a sufficient condition for DSC is that $V_t$ is a first-order Markov chain with a stochastically increasing transition distribution. This is shown in the following proposition, which uses a result from the literature on stochastic orders (see Appendix E for a proof).

**Proposition 1.** *Suppose that $U_t(y)$ is determined by (5). If $V_t$ is a first-order Markov chain (conditional on $A, Y_0$) and $\mathbb{P}[V_{t+1} \leq v_{t+1} | V_t = v_t, A, Y_0]$ is decreasing in $v_t$ for all $v_{t+1}$, then DSC is satisfied.*

If $(V_t, V_{t+1})$ is jointly normally distributed conditional on $A$ and $Y_0$, then the positive stochastic monotonicity condition of Proposition 1 is equivalent to the correlation coefficient between $V_t$ and $V_{t+1}$ (given $A, Y_0$) being non-negative. However, Proposition 1 is sensitive to the inclusion of covariates in (5). If $X$ is time-invariant, then all of the conditions can be modified to be conditional on $X$. On the other hand, if $X$ is time-varying then changes in the index $X_t'\beta$ would need to be accounted for to determine whether DSC can be justified. In particular, if $X$ is strictly exogenous in the sense of A1, and if the index $X_t'\beta$ also satisfies a stochastic monotonicity condition like that on $V_t$, then a modification of Proposition 1 can still be shown to hold.[20] Lower-level conditions for the index $X_t'\beta$ (vs. any given component of $X_t$) to satisfy this property are more difficult to motivate.

Column (7) of Table 2S reports identified sets when DSC is added to MTR and ST (with $m = 2$). The only parameter that is substantially affected is $\text{SD}_t^+(\cdot|000)$, which has a slightly larger lower bound under DSC. In other applications, DSC might have more content.

#### 4.3.4 Monotone Instrumental Variables

The role of covariates in a partial identification analysis is often markedly different than in analyses premised on point identification. In the DBR model of Section 2, conditioning on a richer set of covariates is often viewed as a way to make A1 and

[19] If ST does not hold, then the statement that $\mathbb{P}_P[U_t(y) = 1, U_{t+s}(y) = 1]$ is decreasing in $|s|$ is equivalent to the statement that $(U_t(y), U_{t+s}(y))$ is decreasing in the upper orthant order with respect to $|s|$, see e.g. Shaked and Shanthikumar (2007, Section 6.G). However, the upper orthant order does not necessarily have a clear interpretation as a positive dependence concept.

[20] This statement can be justified by applying Theorem 9.A.1. of Shaked and Shanthikumar (2007), which shows that the concordance ordering is closed under convolution.

A2 more likely to hold, both because adding covariates "removes" them from the latent variables $A$ and $V_t$, and because conditioning on more covariates may make it more likely that $V_t$ and $A$ are conditionally independent in A1. In contrast, in the DPO model, there is no parametric index structure through which the roles of the latent variables $U_t(y)$ and $X$ are made comparable. Moreover, additional maintained assumptions such as ST are made stronger by conditioning on covariates, not weaker.

If an analyst believes that a particular covariate is statistically independent of the potential outcomes, then this can be added as an exclusion (instrumental variable) restriction. Such a restriction can be either conditional on other covariates, or unconditional. A weaker form of an instrumental variable condition only assumes the direction of the dependence between the latent variables and the proposed instrument. This is the monotone instrumental variable (MIV) assumption introduced by Manski and Pepper (2000, 2009). In the context of the dynamic potential outcomes model, an MIV-type assumption is as follows.

***Assumption MIV:*** *Every $P \in \mathcal{P}^{\dagger}$ is such that $\mathbb{P}_P[U_t(y) = 1|X = x]$ is weakly increasing in one or more component of $x$ for $y = 0, 1$ and every $t \geq 1$.*

Of course, one can strengthen MIV to be a full exclusion restriction by imposing both directions of weak monotonicity with respect to the same covariate component.

For the current application, I consider the following MIV assumption:

$$\begin{aligned} &\mathbb{P}_P[U_t(y) = 1|\text{education} = e, \text{husband's income} = h] \\ &\quad \text{is increasing in } e \text{ and decreasing in } h \text{ for } y = 0, 1 \text{ and all supported } (e, h), \end{aligned} \tag{14}$$

where education is the woman's maximum attained education level (less than 12 years, 12 years, 12–16 years, or more than 16 years) and husband's income is the quintile of the measure of permanent income defined previously. The MIV condition in (14) assumes that education has a positive effect on labor force participation, conditional on husband's income. This seems reasonable if higher education increases the opportunity cost of non-participation. Second, the condition assumes that permanent income has a negative effect on labor force participation, conditional on education. This seems reasonable if marginal utility is decreasing in overall household income.

Observe that in the parametric DBR model,

$$\mathbb{P}[U_t(y) = 1|X = x] = \mathbb{E}\left[\Phi(\gamma y + x_t'\beta + A)\right], \tag{15}$$

where $x_t$ is the subcomponent of $x$ that is included in the conditioning set at time $t$ and $\Phi$ is the standard normal distribution function. Hence, the parametric DBR

model implies that $\mathbb{P}[U_t(y) = 1|X = x]$ is monotone increasing in a component of $x$ if the sign of the corresponding $\beta$ component is positive. In such a model, MIV amounts to placing a priori sign restrictions on components of covariate coefficients $\beta$. Imposing MIV in both directions—i.e. assuming that a particular component of $\beta$ is 0—corresponds to an exclusion restriction. Exclusion restrictions like these are often imposed in applications of the parametric DBR by not including various leads and lags of the time-varying components of $X$.

Column (8) of Table 2S reports identified sets for specifications that maintain the above MIV assumption along with MTR and ST with $m = 2$. For most parameters MIV causes a modest decrease in the upper bound. However, under the MTS assumption discussed ahead, MIV is shown to have no additional identifying content. The final nonparametric specification will therefore not impose the MIV condition, since doing so has no practical benefit in the current application. Additional MIV assumptions will be considered in a semiparametric framework that is currently under development.

#### 4.3.5 Monotone Treatment Selection

For the static potential outcomes model, Manski and Pepper (2000) considered the identifying content of assuming that potential outcomes are larger for agents who select into treatment than for those who do not. This monotone treatment selection (MTS) condition captures the idea that an analyst may be willing to make a priori assumptions on the direction of bias that would arise from a simple treatment–control contrast of an endogenously assigned treatment. For the DPO model, the distinction between treatment and control is less stark than in the static case, since each of an agent's past outcomes could be viewed as a "treatment." Hence, there appear to be a number of different MTS-like conditions one could reasonably consider. I will focus on the following variant.

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

$$\mathbb{P}_P[U_t(y) = 1|Y_{t-1} = 1, Y_{t-2} = y_{t-2}] \geq \mathbb{P}_P[U_t(y) = 1|Y_{t-1} = 0, Y_{t-2} = y_{t-2}] \quad (16)$$

*for $y = 0, 1$, $y_{t-2} = 0, 1$ and all $t \geq 2$.*

The $y = 0$ part of MTS says that women who were employed in year $t - 1$ are more likely to be employed in year $t$, even in the counterfactual state that they were actually non-employed in year $t - 1$. Similarly, the condition with $y = 1$ says that women who were non-employed in year $t - 1$ are less likely to be employed in year $t$, even if they had (counterfactually) been employed in period $t - 1$. Hence, MTS can

be interpreted as saying that women who were employed in period $t-1$ have a higher latent propensity to be employed in period $t$ than women who were non-employed in period $t-1$, conditional on having the same employment status in period $t-2$.

The additional conditioning on $Y_{t-2} = y_{t-2}$ in these statements ensures that the participation decision in year $t-1$ is comparable. That is, since the event $[Y_{t-1} = y_{t-1}, Y_{t-2} = y_{t-2}]$ is equivalent to the event $[U_{t-1}(y_{t-2}) = y_{t-1}, Y_{t-2} = y_{t-2}]$, conditioning on $Y_{t-2} = y_{t-2}$ ensures that the conditioning events on the left and right sides of (16) are expressed in terms of the same potential outcome $U_{t-1}(y_{t-2})$. A stronger form of MTS would extend this conditioning all the way back to $t = 0$. I found that this contributed no additional identifying content while imposing additional computational burden, due to the large number of additional restrictions.

To evaluate the credibility of MTS, consider when (16) would be satisfied in the parametric DBR model. There, (16) becomes (suppressing covariates)

$$\begin{aligned} &\mathbb{P}\left[A + V_t \geq -\gamma y | A + V_{t-1} \geq -\gamma y_{t-2}, Y_{t-2} = y_{t-2}\right] \\ &\quad \geq \mathbb{P}\left[A + V_t \geq -\gamma y | A + V_{t-1} < -\gamma y_{t-2}, Y_{t-2} = y_{t-2}\right]. \end{aligned} \tag{17}$$

A sufficient condition for (17) is that $A + V_t$ is positive quadrant dependent with $A + V_{t-1}$, conditional on $Y_{t-2}$.[21] If $V_t$ and $A$ are independent and normally distributed, as they are under A1 and A2, then this is satisfied if and only if the correlation between $A + V_t$ and $A + V_{t-1}$ is positive, i.e. $\mathbb{V}(A) \geq -\operatorname{Cov}(V_t, V_{t-1})$. If the index in the parametric DBR is meant to represent a reduced form for the net utility (broadly defined) of employment, then one would typically expect transitory shocks $V_t$ to wages or private information to not be strongly negatively correlated, in which case (16) would be satisfied. Viewed through the context of a parametric DBR, MTS therefore appears to be quite reasonable.

Column (9) of Table 2S reports identified sets when MTS is added to MTR, ST (with $m = 1$), DSC and the MIV condition given in the previous section. Comparing columns (8) and (9), it is clear that MTS has a dramatic effect on the upper bounds of positive state dependence parameters.

#### 4.3.6 Fixed Effects or "Time is an Instrument"

The parametric DBR model maintains a distinction between time-invariant latent factors $A$ and time-varying latent factors $V_t$. Distinctions like this are also imposed in the

[21] Two random variables $A$ and $B$ are positive quadrant dependent if $\mathbb{P}[A > a, B > b] \geq \mathbb{P}[A > a]\,\mathbb{P}[B > b]$ for every $a$ and $b$, see e.g. Nelsen (2006). It is straightforward to show that if $A$ and $B$ are continuously distributed then this implies that $\mathbb{P}[A \geq a | B \geq b] \geq \mathbb{P}[A \geq a | B < b]$.

nonparametric literature surveyed in Section 2. In that literature, researchers have used the difference between $A$ and $V_t$ to assume that $Y_t$ follows a homogenous, first-order Markov process, conditional on $A$ and the initial period $Y_0$. This essentially treats $A$ as a time-invariant fixed effect, in the manner often desired for the parametric DBR model. In particular, Chernozhukov et al. (2013) derived non-sharp analytic bounds on the ATE for the nonparametric model (3) under the assumption that $V_t|Y_{t-1}, \ldots, Y_0, A$ is distributed like $V_1|Y_0, A$ for every $t \geq 1$. They described this condition as assuming that "time is randomly assigned" or "time in an instrument" (TIV).

***Assumption TIV.*** *For every $P \in \mathcal{P}^\dagger$, there exists a random variable $A$ such that the distribution of $U_t(y)|Y_{t-1}, \ldots, Y_0, A$ under $P$ is the same as the distribution of $U_1(y)|Y_0, A$ under $P$ for $y = 0, 1$ and all $t$.*

Point estimates of the Chernozhukov et al. (2013) bounds are given in column (C1) of Table 2S. Compared to the identified set in column (9), the bounds in column (C1) are quite wide and relatively uninformative. Column (C2) shows the slightly improved non-sharp lower bound that can be obtained from the results of Chernozhukov et al. (2013) by imposing MTR, in which case $\text{ATE}_t = \text{SD}_t^+$. As the next proposition shows, it is also possible to impose the TIV assumption in the DPO model. Moreover, unlike for the nonparametric model of (3), Theorem 1 allows one to combine this assumption with the other assumptions already discussed.

**Proposition 2.** *If TIV holds then for every $P \in \mathcal{P}^\star$, $y = 0, 1$, and all $t \geq 1$,*

$$\mathbb{P}_P[U_t(y) = 1] \geq \sum_{t=0}^{T-1} \mathbb{P}[Y_t = 1, Y_{t-1} = y, Y_s \neq y \; \forall s < t-1] \quad \textit{and}$$

$$\mathbb{P}_P[U_t(y) = 1] \leq \sum_{t=0}^{T-1} \mathbb{P}[Y_t = 1, Y_{t-1} = y, Y_s \neq y \; \forall s < t-1] + \mathbb{P}[Y_s \neq y \; \forall s \leq T-1].$$

Proposition 2 is simply a restatement of a result of Chernozhukov et al. (2013) put in terms of the dynamic potential outcomes model.[22] The implied bounds in Proposition 2 are linear in $P = \{P[u, x] : u \in \mathcal{U}, x \in \mathcal{X}\}$, and so can be imposed with Theorem 1 together with any of the other assumptions an analyst wishes to maintain. However, since Proposition 2 only establishes an implication of TIV, and not an equivalence, the resulting identified set may also include structures that do not satisfy TIV. Hence, as in Chernozhukov et al. (2013), an identified set constructed by imposing the bounds in Proposition 2 may be non-sharp.

[22]For completeness, a proof of Proposition 2 is provided in Appendix F.

Column (10) of Table 2S reports non-sharp identified sets that add the restrictions in Proposition 2 to MTR, ST (with $m = 2$), DSC, MIV, and MTS. Evidently, there is no additional identifying content contained in TIV. It is unclear whether this is due to the non-sharpness of the identified set, or because TIV actually contains no identifying content beyond the other maintained assumptions in the current application.

Regardless of its identifying content, TIV has some strong and undesirable implications in the context of female labor force participation. Foremost among these is its implication that the fixed effect $A$ is the only source of serial correlation in latent variables. Hyslop (1999) found strong evidence against this assumption in a parametric DBR model, instead concluding that the $V_t$ terms in (1) were more appropriately modeled as an AR(1) process. If $V_t$ follows an AR(1) process, then $U_t(y)$ (as implied by that model, i.e. (5)) will be dependent with all past lags of $Y_t$, even conditional on $A$ and hence TIV will not be satisfied.[23] Serial correlation in the time-varying unobservables seems likely, since they will represent temporal changes in productivity, human capital and private information. This problem is compounded in a nonparametric setting, since any serially correlated observed time-varying covariates that are omitted will end up being reflected in the time-varying latent term.

Given these objections and the lack of identifying content contributed by TIV, it seems just as well to remove the assumption. Similarly, after some experimentation, I found that MIV also did not add any additional identifying content once MTS was added, so it is removed as well. Column (11) of Table 2S confirms that the resulting identified sets (maintaining only MTR, ST with $m = 2$, DSC and MTS) are unchanged.

## 5 Statistical Inference

The identified sets $\Theta^\star$ constructed in the previous section treat the empirical distribution of the observed data as if it were the population distribution. In this section, I discuss the construction of confidence sets for $\Theta^\star$ that account for the sampling variation that arises when viewing the empirical distribution as resulting from an i.i.d. sample from some underlying population distribution. These confidence sets contain (with probability at least $1 - \alpha$) the parameter $\theta_0 = \theta(P_0) \in \Theta^\star$ corresponding to the "true" structure $P_0 \in \mathcal{P}^\star$ that generated the data.

[23] This point was raised early in the literature by Heckman (1981a) and Chamberlain (1984).

### 5.1 Construction of Confidence Intervals

Some additional notation is required. Let $\mathcal{W} \equiv \text{supp}(Y, X)$ denote the joint support of the observable data $W \equiv (Y, X)$. For each $w \equiv (w_y, w_x) \in \mathcal{W}$ define

$$m_{\text{oeq},w}(W, P) \equiv \mathbb{1}[Y = w_y, X = w_x] - \sum_{u \in \mathcal{U}_{\text{oeq}}(w_y)} P[u, w_x], \tag{18}$$

Next, partition the restriction function $\rho$ into a deterministic component $\rho_d$, and a stochastic component $\rho_s$ with dimension $d_s$. The deterministic component, $\rho_d : \mathcal{P} \to \mathbb{R}^{d_\rho - d_s}$, is a function defined on $\mathcal{P}$ alone that does not depend on the distribution of $W$. The stochastic component, $\rho_s : \mathcal{P} \times \mathcal{F}_W \to \mathbb{R}^{d_s}$, is a function defined on $\mathcal{P}$ and the collection $\mathcal{F}_W$ of possible distributions for $W$. Furthermore, assume that $\rho_s$ can be represented as a moment condition, i.e. that there exists a function $m_\rho : \mathcal{W} \times \mathcal{P} \to \mathbb{R}^{d_s}$ that is linear in $P$ for fixed $F$, and for which $\rho_s(P, F) = \mathbb{E}_F[m_\rho(W, P)]$, where $F \in \mathcal{F}_W$ and $\mathbb{E}_F$ is the expectation computed when $W$ is distributed as $F$. This condition is satisfied by all of the auxiliary identifying assumptions utilized in Section 4.

For example, suppose that there are no covariates so that $\mathcal{X}$ can be taken as a singleton and $W = Y$. Then MTR would be represented through $\rho_d$, since it is not a restriction that depends on the distribution of observables $W$. On the other hand, MTS would be part of $\rho_s$, since it depends on the distribution of $(Y_{t-1}, Y_{t-2})$. With covariates, MIV would be part of $\rho_s$, since it depends on the marginal distribution of $X$. Note that the current notation is not flexible enough to account for components of $\rho_s$ that correspond to equality (vs. inequality) constraints. This is simply because none of the proposed identifying assumptions in Section 4.3 can be classified as such. Accommodating such restrictions is immediate, but requires some additional notation.

Next, define $\mathcal{P}_d^\dagger \equiv \{P \in \mathcal{P} : \rho_d(P) \geq 0\}$ as the set of constraints on $P$ that do not depend on the distribution of $W$. These include not only $\rho_d$, 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

$$\begin{aligned} \mathcal{P}^\star = \{P \in \mathcal{P}_d^\dagger : \mathbb{E}[m_{\text{oeq},w}(W, P)] = 0 \; \forall w \in \mathcal{W} \\ \text{and } \; \mathbb{E}[m_{\rho,s}(W, P)] \geq 0 \; \forall s = 1, \ldots, d_s\}, \end{aligned} \tag{19}$$

where $m_{\rho,s}(W, P)$ denotes the $s^{\text{th}}$ component of $m_\rho(W, P)$. Equation (19) shows that the DPO model can be viewed as a moment inequality model with parameter space $\mathcal{P}_d^\dagger$, moment equalities $\{\mathbb{E}[m_{\text{oeq},w}(W, P)] = 0\}_{w \in \mathcal{W}}$, and moment inequalities $\{\mathbb{E}[m_{\rho,s}(W, P)] \geq 0\}_{s=1}^{d_s}$. This observation was not helpful for computing identified sets, and in fact would have obscured the ease with which $\mathcal{P}^\star$ can be computed given

knowledge of the distribution of $W$. On the other hand, the moment inequality framework is quite useful for inference, since several recent papers have carefully considered the delicate theoretical issues involved in moment inequality models.[24]

In particular, I will construct confidence regions using the approach of Bugni et al. (2014) (BCS). The BCS approach adapts the generalized moment selection (GMS) approach of Andrews and Soares (2010) (AS) to a form that is more amenable to conducting inference on low-dimensional parameters.[25] The immediate way to use GMS to construct a confidence region for a low-dimensional parameter of interest $\theta_0 \equiv \theta(P_0)$ is to first construct a confidence region for $P_0$ through test-inversion, and then form the image of this confidence region under $\theta$. BCS argue that this can be quite conservative. But more importantly, the first step of this procedure is computationally infeasible in the DPO model, since it requires performing a grid-search on a space of dimension $|\mathcal{U}| \times |\mathcal{X}| = 2^{2T+1} \times |\mathcal{X}|$. The BCS approach effectively replaces this high-dimensional grid search with a series of optimization problems.

One implementation of the BCS procedure is as follows. For notational convenience, 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 $m_j$ correspond to $\{m_{\rho,s}\}_{s=1}^{d_s}$. Quantify the moment inequalities and equalities with the criterion function

$$Q(P) \equiv \sum_{j=1}^{d_s} \left[ \frac{\mathbb{E}[m_j(W,P)]}{\mathbb{V}[m_j(W,P)]^{1/2}} \right]_-^2 + \sum_{j=d_s+1}^{d_m} \left[ \frac{\mathbb{E}[m_j(W,P)]}{\mathbb{V}[m_j(W,P)]^{1/2}} \right]^2,$$

where $[x]_- \equiv \min\{0, x\}$ and $\mathbb{V}$ denotes the variance operator under the population distribution of $W$. Notice that for $P \in \mathcal{P}_d^\dagger$, $Q(P) = 0$ if and only if $P \in \mathcal{P}^\star$. This criterion is referred to in the literature as the modified method of moments. It is attractive for its computational ease, but other criterion functions are possible.[26] Given

[24] It is not clear how one could construct valid confidence regions by using empirical analogs of the programs in Theorem 1. Andrews and Han (2009) show that naively bootstrapping (or subsampling) empirical analogs of $\theta_l^\star$ and $\theta_u^\star$ will not lead to consistent confidence regions. Freyberger and Horowitz (2013) analyze an instrumental variables model for which the identified set can also be represented through the solution to two linear programming problems, although their programs have more structure than those considered here. They propose a modified bootstrap procedure based on the sample analogs of the solutions to the linear programs. Importantly, their procedure accounts for the discontinuity in the asymptotic distribution of these solutions by restricting attention to the set of nearly–optimal basic solutions. In the DPO model, the set of basic solutions is extremely large, so a similar procedure would be computationally difficult and might have poor finite sample properties. Their procedure is also premised on the assumption that the identified set computed with the empirical distribution is non-empty, which is not necessarily desirable for analyzing the DPO model.

[25] See also Gandhi et al. (2013) (Section 5) who consider an inferential approach that is effectively subsumed by the procedure later proposed by Bugni et al. (2014).

[26] In particular, criterion functions that incorporate information on cross-moment correlations may be

an i.i.d. sample $\{W_i\}_{i=1}^n$ of size $n$, a sample analog of $Q$ is constructed by replacing $\mathbb{E}$ and $\mathbb{V}$ with their empirical counterparts:

$$Q_n(P) \equiv \sum_{j=1}^{d_s} \left[\frac{\sqrt{n}\overline{m}_{n,j}(P)}{\sigma_{n,j}(P)}\right]_-^2 + \sum_{j=d_s+1}^{d_m} \left[\frac{\sqrt{n}\overline{m}_{n,j}(P)}{\sigma_{n,j}(P)}\right]^2,$$

$$\text{where } \overline{m}_{n,j}(P) \equiv \frac{1}{n}\sum_{i=1}^n m_j(W_i, P), \qquad \sigma_{n,j}^2(P) \equiv \frac{1}{n}\sum_{i=1}^n (m_j(W_i, P) - \overline{m}_{n,j}(P))^2,$$

and scaling by the rate of convergence has already been applied.

Given a parameter $\theta$, a profiled version of $Q_n$ can be defined as

$$\overline{Q}_n(t) \equiv \inf_{P \in \mathcal{P}_d^\dagger(t)} Q_n(P) - \inf_{P \in \mathcal{P}_d^\dagger} Q_n(P),$$

where $\mathcal{P}_d^\dagger(t) \equiv \{P \in \mathcal{P}_d^\dagger : \theta(P) = t\}$ and the objective function has been re-centered around $\inf_{P \in \mathcal{P}_d^\dagger} Q_n(P)$ to improve power (Chernozhukov et al., 2007).[27] The profiled criterion $\overline{Q}_n(t)$ will serve as a test statistic for a test of the null hypothesis $H_0 : t \in \Theta^\star$. Confidence regions for $\Theta^\star$ can be constructed through test inversion, i.e. by collecting all $t \in \Theta$ for which this null hypothesis is not rejected.

To operationalize these tests, one needs a way to approximate the distribution of $\overline{Q}_n(t)$ so as to construct an appropriate critical value. This is difficult because the asymptotic distribution of $Q_n(P)$ is discontinuous in the number and identity of the inequality moments that bind, i.e. those $j \in \{1, \ldots, d_s\}$ for which $\mathbb{E}[m_j(W, P)] = 0$. Inequality moments that do not bind do not affect the asymptotic distribution of $Q_n(P)$ for a fixed $F \in \mathcal{F}_W$, so removing these moments from (or limiting their effect on) a proposed approximation is important for achieving inference that is not excessively conservative. The GMS procedure introduced by AS solves this problem by effectively smoothing out the discontinuity in the asymptotic distribution of $Q_n(P)$. This is accomplished by approximating the limiting distribution of $Q_n(P)$ under a drifting sequence of $F \in \mathcal{F}_W$ in such a way that the limit accounts for the distance of the inequality moments from 0 in the population.[28]

In particular, one implementation of the GMS procedure considers the asymptotic

preferable (see AS and Andrews and Barwick (2012)) but are more difficult to compute.

[27] Note that here I am assuming that $\theta$ does not depend on $F_W$. This rules out parameters that condition on past outcomes such as $\text{SD}_t^+(\cdot|0)$.

[28] My understanding is that similar ideas were independently and contemporaneously derived by several other authors including Bugni (2010) and Canay (2010)—see AS for details on the related literature.

distribution of

$$Q_n^{\text{GMS}}(P) \equiv \sum_{j=1}^{d_s} \left[\nu_{n,j}^*(P) + \xi_{n,j}(P)\right]_-^2 + \sum_{j=d_s+1}^{d_m} \left[\nu_{n,j}^*(P)\right]^2, \tag{20}$$

$$\text{where } \nu_{n,j}^*(P) \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{m_j(W_i^*, P) - \overline{m}_{n,j}(P)}{\sigma_{n,j}(P)}, \quad \xi_{n,j}(P) \equiv \kappa_n^{-1} \frac{\sqrt{n}\overline{m}_{n,j}(P)}{\sigma_{n,j}(P)},$$

$\{W_i^*\}_{i=1}^n$ is a bootstrap sample drawn i.i.d. with replacement from $\{W_i\}_{i=1}^n$, and $\kappa_n$ is a tuning parameter that I will always set to $\log(n)^{1/2}$ as recommended by AS. Under this specification the contribution of the $j^{\text{th}}$ moment inequality to $Q_n^{\text{GMS}}(P)$ increases smoothly as the $j^{\text{th}}$ moment inequality becomes more strongly violated and $\overline{m}_{n,j}(P)$ becomes more negative.[29] The asymptotic distribution of $Q_n^{\text{GMS}}(P)$ can be approximated through simulation by redrawing $\{W_i^*\}_{i=1}^n$ a large number (say, $B$) of times. AS show that a test that rejects $H_0 : P \in \mathcal{P}^\star$ whenever $Q_n(P)$ is larger than the simulated $1-\alpha$ quantile of $Q_n^{\text{GMS}}(P)$ will have asymptotic size $\alpha$, while also having good power and uniformity properties.

The GMS procedure is designed for constructing confidence regions for $\mathcal{P}^\star$. Projecting these confidence regions for $\mathcal{P}^\star$ to obtain confidence regions for $\Theta^\star$ is computationally infeasible if $P$ is high-dimensional. Moreover, BCS show that directly simulating the profiled GMS criterion function $\inf_{P \in \mathcal{P}_d^\dagger(t)} Q_n^{\text{GMS}}(P)$ does not produce a test that controls size asymptotically. Instead, BCS propose two GMS–like procedures for approximating the asymptotic distribution of $\overline{Q}_n(t)$. The first procedure simulates the distribution of

$$\overline{Q}_n^{\text{R1}}(t) \equiv Q_n^{\text{GMS}}(\widehat{P}_n(t))$$

where $\widehat{P}_n(t) \in \arg\inf_{P \in \mathcal{P}_d^\dagger(t)} Q_n(P)$ is a minimizer of $Q_n(P)$ over $\mathcal{P}_d^\dagger(t)$.[30,31] The second

[29] AS discuss several possible ways to account for $\xi_{n,r}(P)$ in $Q_n^{\text{GMS}}(P)$. The form chosen in (20) corresponds to their "$\varphi^{(4)}$" GMS function. This choice is convenient computationally, since it preserves smoothness and convexity of $Q_n$, as discussed ahead. However, Andrews and Barwick (2012) present simulation evidence that suggests non-smooth choices may perform better in finite samples.

[30] Note that $Q_n(P)$ is convex but not necessarily strictly convex. Therefore, its minimizer over a convex set may not be unique. BCS specify $\overline{Q}_n^{\text{R1}}(t) \equiv \inf_{P \in \widehat{\mathcal{P}}_n^\star(t)} Q_n^{\text{GMS}}(P) - Q_n^{\text{rc}}$ where $\widehat{\mathcal{P}}_n^\star(t) = \arg\inf_{P \in \mathcal{P}_d^\dagger(t)} Q_n(P)$ is the collection of all minimizers of $Q_n(P)$. For computational considerations, I replace this infimum by a single point. This still yields a test that controls size, although it will be more conservative.

[31] Recentering R1 does not seem to be appropriate. For example, if $v_{n,j}^*(P)$ does not depend on $P$ (which often happens, see below) and there are no moment inequalities, then $Q_n^{\text{GMS}}(P)$ does not depend on $P$ and so a recentered version of $\overline{Q}_n^{\text{R1}}(t)$ would be deterministically 0.

procedure simulates the distribution of

$$\overline{Q}_n^{\mathrm{R2}}(t) \equiv \inf_{P \in \mathcal{P}_d^\dagger(t)} Q_n^{\mathrm{R2}}(P) - \inf_{P \in \mathcal{P}_d^\dagger} Q_n^{\mathrm{R2}}(P),$$

$$\text{where} \quad Q_n^{\mathrm{R2}}(P) \equiv \sum_{j=1}^{d_s} \left[\nu_{n,j}^*(P) + \xi_{n,j}(P)\right]_-^2 + \sum_{j=d_s+1}^{d_w} \left[\nu_{n,j}^*(P) + \xi_{n,j}(P)\right]^2, \qquad (21)$$

and $\inf_{P \in \mathcal{P}_d^\dagger} Q_n^{\mathrm{R2}}(P)$ accounts for re-centering. BCS show that a test that rejects $H_0 : t \in \Theta^\star$ when $\overline{Q}_n(t)$ is larger than the simulated $1-\alpha$ quantile of either $\overline{Q}_n^{\mathrm{R1}}(t)$ or $\overline{Q}_n^{\mathrm{R2}}(t)$ will control size asymptotically and have good uniformity properties. They also propose a third test called the "minimum resampling" (MR) test that rejects $H_0 : t \in \Theta^\star$ when $\overline{Q}_n(t)$ is larger than the simulated $1-\alpha$ quantile of

$$\overline{Q}_n^{\mathrm{MR}}(t) \equiv \min\{\overline{Q}_n^{\mathrm{R1}}(t), \overline{Q}_n^{\mathrm{R2}}(t)\}.$$

BCS establish that the MR test also controls size asymptotically. In addition, since its critical value is by construction smaller than those for tests based on R1 or R2 alone, the MR test has at least weakly better power than tests based on either R1 or R2. In the following, I refer to the test that rejects $H_0 : t \in \Theta^\star$ when $\overline{Q}_n(t)$ is larger than the simulated $1-\alpha$ quantile of $\overline{Q}_n^{\mathrm{MR}}(t)$ as the MR test. A $1-\alpha$ MR confidence region for $\Theta^\star$ is the set of all $t$ for which the MR test does not reject.

As an alternative to the MR test, I also consider confidence regions constructed through subsampling (Chernozhukov et al., 2007; Romano and Shaikh, 2008, 2010). The subsampling procedure is more straightforward than a GMS–based approach, but BCS argue that it can have worse asymptotic power than the MR test.[32] The Monte Carlo simulation in the next section suggests that, at least for the DPO model and with finite samples, the relative power ranking of the SS and MR tests depends on the null hypothesis, i.e. on $t$. The subsampling approach approximates the asymptotic distribution of $\overline{Q}_n(t)$ by the distribution of

$$\overline{Q}_{b_n}^{\mathrm{SS}}(t) \equiv \inf_{P \in \mathcal{P}_d^\dagger(t)} Q_{b_n}^{\mathrm{SS}}(P) - \inf_{P \in \mathcal{P}_d^\dagger} Q_{b_n}^{\mathrm{SS}}(P),$$

where $Q_{b_n}^{\mathrm{SS}}(P)$ is analogous to $Q_n(P)$, but constructed instead using a randomly drawn subsample (without replacement) $\{W_i^*\}_{i=1}^{b_n}$ of size $b_n$ from $\{W_i\}_{i=1}^n$. This profiled subsampling procedure was first suggested in Romano and Shaikh (2008). In the following, I refer to the test that rejects $H_0 : t \in \Theta^\star$ when $\overline{Q}_n(t)$ is larger than the

[32] Similar results are established by AS for non-profiled GMS-based tests.

$1-\alpha$ quantile of $\overline{Q}_{b_n}^{\mathrm{SS}}(t)$ 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 reported ahead in Section 5.3 suggest that both the MR and SS tests have poor finite-sample power in the DPO model. As a result, confidence regions constructed by inverting these tests tend to be excessively wide. As a simple way to ameliorate this problem, I consider a third test, referred to as the "minimum quantile" (MQ) test, with critical value taken as the minimum of the critical values for the MR and SS tests. The MQ test will be asymptotically the same as either the MR or SS test for a given null hypothesis $H_0 : t \in \Theta^\star$. For constructing confidence regions, however, the MQ test will differ from the MR and SS tests to the extent that the power rankings of the MR and SS tests vary with $t$ and with the sample.

A fourth test I consider is the "quantile of the minimum" (QM) test, which rejects $H_0 : t \in \Theta^\star$ if $\overline{Q}_n(t)$ is larger than the $1-\alpha$ simulated/subsampled quantile of $\min\{\overline{Q}_n^{\mathrm{MR}}(t), \overline{Q}_{b_n}^{\mathrm{SS}}(t)\}$. However, it is not clear that this test controls size. It may be possible to adapt the results on size control for the MR test given in Bugni et al. (2014) to the QM test, since the resampling statistics (R1 and R2) and the subsampling statistic have asymptotic distributions that have a common structure, however this has not been shown. The Monte Carlo results in Section 5.3 suggest that the QM test is considerably more powerful than the other three tests considered. However, since it is not clear that it controls size, I do not use the QM test when constructing confidence intervals for the PSID data.

## 5.2 Computational Considerations

In order to implement the tests in the previous section, it is important to be able to reliably solve the optimization problems that define $\overline{Q}_n(t)$, $\overline{Q}_n^{\mathrm{R1}}(t)$, $\overline{Q}_n^{\mathrm{R2}}(t)$, and $\overline{Q}_n^{\mathrm{SS}}(t)$. Reliability—specifically, 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 bootstrapping/subsampling and inverting hypothesis tests into confidence intervals. All of these problems are convex programs if (i) $\mathcal{P}_d^\dagger(t)$ is determined by the intersection of linear equalities and inequalities; (ii) $\overline{m}_{n,j}(P)$ is linear in $P$ for all $j$; (iii) $\sigma_{n,j}(P)$ does not depend on $P$; and (iv) $\xi_{n,j}(P)$ enters the GMS objective function linearly, as in (20). Convex programs of this sort are relatively easy to solve quickly and reliably.[33] Conditions (i) and (ii) are satisfied for all combinations of the parameters and auxiliary identifying assumptions that were discussed in Section

[33]Note that even under conditions (i)–(iv), finding $\overline{Q}_n(t)$ is not necessarily a quadratic program, because of the $[\cdot]_-$ function for the moment inequalities. In situations where there are no moment inequalities, the extra quadratic structure can be exploited.

4. Condition (iv) corresponds to a particular choice of the GMS function ("$\varphi$" equal to $\varphi^{(4)}$) in AS.

However, condition (iii) effectively requires $m_j(W, P)$ to be additively separable in $P$. Separability holds for the observational equivalence moments (18), but not for some stochastic constraints, such as MTS. Including such nonseparable moments would generally make $\overline{m}_{n,j}(P)/\sigma_{n,j}(P)$ a nonlinear function of $P$. As a result, determining $\overline{Q}_n(t)$ would require solving a high-dimensional optimization problem with a convex constraint set but a potentially non-convex objective function. This can be quite difficult. To avoid this problem, but still include nonseparable constraints like MTS, I modify the definition of $Q$ (and $Q_n$) so that these nonseparable moments are not scaled by their standard deviations (or sample standard deviations). This restores convexity in the objective function at the cost of losing the scale invariance property of the criterion function.[34]

### 5.3 Monte Carlo Simulation

This section discusses the results of a small Monte Carlo study aimed at gauging the finite sample performance of the previously discussed tests when applied to the DPO model. The data generating process is taken to be the empirical distribution observed in the PSID data. In order to moderate the computational requirements, I use only the first five periods of data ($T = 4$) and keep only the subset of observations that experience two or fewer transitions (i.e. instances in which $Y_t \neq Y_{t-1}$) over this period. This leaves $n =$ 1,735 out of the 1,812 original cross-sectional observations. The empirical distribution of $Y$ for this subpopulation is reported in Table 3. The data generating process in the Monte Carlo simulation draws $Y$ according to this empirical distribution.

Tables 4 and 5 report features of the distribution of estimated upper and lower bounds, as well as the rejection rates of the MR, SS, MQ and QM tests of $H_0 : t \in \Theta^\star$ at nominal level $\alpha = .05$ for several values of $t$.[35] The first specification, reported in Table 4, imposes MTR and ST with $m = 2$, while the second specification, reported in Table 5, imposes MTR, ST with $m = 2$, and MTS. In both cases the parameter is $\text{SD}_t^+$.[36] The first specification represents a case in which there are no moment inequalities and the criterion function is scale-invariant, while in the second specification, the addition of

[34]Some of the earlier papers on moment inequalities, such as Chernozhukov et al. (2007), Romano and Shaikh (2008) and Ciliberto and Tamer (2009) considered modified methods of moments estimators that are not scale invariant. AS argue that this may lead to poor power.

[35]The simulations were conducted with 500 replications, $B = 500$, $\kappa_n = \sqrt{\log(n)}$, and $b_n = n^{2/3}$. These choices of $\kappa_n$ and $b_n$ were recommended by AS and Bugni (2010), respectively.

[36]Note that the true values listed in Table 4 differ from those in Table 2S because here $T = 4$.

MTS requires moment inequalities that are not scale-invariant due to the computational considerations discussed in the previous section.

The results suggest that the MR, SS and MQ tests are highly conservative, as all three tests only seldom reject the null hypothesis for values of $t$ (indicated in the table by boxes) at which it is true. All three tests have low power and hence will lead to wide confidence intervals. For the smaller sample size ($n$ = 1,735), the MR test has relatively higher power for values of $t$ on the left side of the identified set, while the SS test has higher power for values of $t$ on the right side of the identified set, but only in the first specification. The MQ test captures both of these areas of relatively good power, producing an overall more powerful test. However, the MQ test is still quite conservative and not very powerful. When $n$ is doubled, the power differences between the tests appear to diminish, although the MR test is still not very powerful to the right of the identified set in the first specification, while the SS test is still not very powerful to the right of the identified set in the second specification. Each of the MR, SS and MQ tests remain highly conservative even with $n$ = 3,470. The QM test is also conservative, but much less so, and has substantially higher power across all hypotheses, both specifications and both sample sizes. However, since it is not known whether the QM test controls size, this could just be an artifact of the specific simulation.

The low power of the MR, SS and MQ tests does not appear to be due to sampling noise in the identified set. If this were the case, one would expect to see substantial sampling variation in the directly computed lower and upper bounds. The simulations provide strong evidence against this explanation. For example, in the first specification (Table 4) with $n$ = 1,735, the MQ test of $H_0 : .04 \in \Theta^\star$ only rejects about 4% of the time, even though .04 is roughly the .01 quantile of the simulated empirical distribution of $\widehat{\theta}_l^\star$. The MQ tests of $H_0 : .02 \in \Theta^\star$ and $H_0 : 0 \in \Theta^\star$ are only rejected 14% and 52% of the time, even though the smallest realization of $\widehat{\theta}_l^\star$ in the simulation was .021. When constructing confidence regions through test inversion, this conservativeness will translate into confidence regions that are excessively wide.

### 5.4 Confidence Intervals for Female Labor Force Participation

In this section I report and discuss 95% confidence intervals for $\text{SD}_t^+$ for three specifications of the DPO model using the full sample of the PSID data.[37] For comparison, recall that the 95% bootstrapped confidence interval of the ATE for the parametric DBR specification discussed at the beginning of Section 4 is [.144, .337]. Here, I con-

[37]The results and discussion in this section are preliminary.

sider three specifications of the DPO model.

The first specification maintains only MTR and ST with $m = 0$. The 95% MQ confidence interval for the ATE (which is equal to $SD_t^+$) in this specification is $[.022, .863]$.[38] While quite wide, this confidence interval is interesting, because it shows that using only a monotonicity and weak stationarity condition, one can reject the hypothesis that there is no state dependence in female labor force participation in a completely nonparametric model. The second specification uses MTR, ST with $m = 2$ and MTS. The 95% MQ confidence interval for the ATE (again equal to $SD_t^+$) in this specification is $[.014, .654]$.

The third specification is like the second specification but increases $m$ from 2 to 4. The identified set for this specification is empty in the sample, but it is still possible to construct a confidence interval by assuming that the model is correctly specified and imposing the various recentering terms discussed in Section 5. The resulting 95% MQ confidence interval for the ATE ($SD_t^+$) is $[.041, .453]$. This is closer to the confidence interval for the parametric DBR model, but obtained under transparent, nonparametric assumptions.

Given the low power of the MQ test exhibited in the simulation studies in the previous section, it is likely that these 95% MQ confidence intervals are highly conservative. Hopefully, future work on inference in partially identified models such as these will provide methods that are less conservative.

## 6 Conclusion

This paper has discussed the use of a dynamic potential outcomes (DPO) model for empirically separating state dependence from unobserved heterogeneity in dynamic binary outcomes. Compared to traditional parametric dynamic binary response (DBR) models, the DPO model has the advantage of being nonparametric, transparent and flexible with regards to its treatment of persistent unobserved heterogeneity. Compared to more recent work on nonparametric DBR models, the DPO model has the advantage of being transparent and amenable to the measurement of many types of parameters under many types and combinations of auxiliary identifying assumptions. It also does not require an analyst to assume that the observed outcomes are conditionally first-order Markov, which may be undesirable for many economic outcomes.

The central challenges of the DPO model are the difficulties with statistical inference raised by the fact the that parameters measuring state dependence are in general

[38] All confidence intervals for the DPO model were constructed using $B = 500$, $\kappa_n = \sqrt{\log(n)}$, and $b_n = n^{2/3}$.

partially, not point, identified. Recent work on inference in moment inequality models has provided a useful starting point for inference in the DPO model. However Monte Carlo simulations suggest that, at least for the DPO model, these methods are highly conservative with low power. It would be helpful to have an inferential method that more directly exploits the linear programming structure of the DPO model. Such a development would have applications beyond the DPO model, but is beyond the scope of the current paper.

Even using tests with poor power, it is possible to use the DPO model to reject at conventional levels the hypothesis of no state dependence in female labor force participation using only weak nonparametric assumptions about stationarity and monotonicity. Under stronger stationarity conditions and an additional assumption of positive dynamic selection into employment (i.e. MTS), I estimate that 4.1–45.3% of married women in the early 1980s were directly affected by state dependence. This confidence interval is consistent with (but substantially wider than) that implied by a standard parametric DBR model.

Table 1: Descriptive Statistics on Labor Force Participation Dynamics

| | time period $t$ | | | | | | |
|---|---|---|---|---|---|---|---|
| | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
| $\mathbb{P}[Y_t = 1]$ | .710 | .694 | .687 | .682 | .700 | .733 | .727 |
| | (.011) | (.011) | (.011) | (.011) | (.011) | (.010) | (.010) |
| $\mathbb{P}[Y_t \neq Y_{t-1}]$ | — | .147 | .143 | .125 | .133 | .124 | .099 |
| | | (.008) | (.008) | (.008) | (.008) | (.008) | (.007) |
| $\mathbb{P}[Y_t = 1 \mid Y_{t-1} = 0]$ | — | .227 | .222 | .194 | .238 | .263 | .174 |
| | | (.015) | (.015) | (.014) | (.014) | (.014) | (.013) |
| $\mathbb{P}[Y_t = 1 \mid Y_{t-1} = 1]$ | — | .885 | .891 | .905 | .916 | .935 | .929 |
| | | (.010) | (.010) | (.009) | (.009) | (.009) | (.008) |
| | total # of transitions | | | | | | |
| | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
| Proportion | .588 | .178 | .146 | .058 | .023 | .007 | .001 |
| of women | (.012) | (.009) | (.008) | (.005) | (.004) | (.002) | — |

Notes: (i) Standard errors in parentheses; (ii) A transition is defined as the event $[Y_t \neq Y_{t-1}]$; (iii) Only one woman experienced 6 transitions so no standard error is provided.

Table 2S: Identified Sets

| | | PDBR | CFHN | | DPO | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | (P) | (C1) | (C2) | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) | (11) |
| MTR | | n/a | | ✓ | | ✓ | | | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| ST $m=0$ | | n/a | n/a | n/a | | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| ST $m=1$ | | n/a | n/a | n/a | | | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| ST $m=2$ | | n/a | n/a | n/a | | | | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| DSC | | n/a | n/a | n/a | | | | | | | ✓ | ✓ | ✓ | ✓ | ✓ |
| MIV | | n/a | n/a | n/a | | | | | | | | ✓ | ✓ | ✓ | |
| MTS | | n/a | n/a | n/a | | | | | | | | | ✓ | ✓ | ✓ |
| TIV | | n/a | ✓ | ✓ | | | | | | | | | | ✓ | |
| $\text{ATE}_t$ | $\theta_l^\star$ | .240 | .098 | .129 | -.126 | .000 | -.066 | -.062 | -.039 | .171 | .171 | .171 | .171 | .171 | .171 |
| | $\theta_u^\star$ | | .716 | .716 | .874 | .874 | .839 | .838 | .824 | .824 | .819 | .717 | .439 | .439 | .439 |
| $\text{SD}_t^+$ | $\theta_l^\star$ | n/a | n/a | .129 | .000 | .000 | .000 | .032 | .058 | .171 | .171 | .171 | .171 | .171 | .171 |
| | $\theta_u^\star$ | | | .716 | .874 | .874 | .853 | .853 | .853 | .824 | .819 | .717 | .439 | .439 | .439 |
| $\text{SD}_t^+(\cdot\|0)$ | $\theta_l^\star$ | n/a | n/a | n/a | .000 | .000 | .000 | .088 | .107 | .326 | .326 | .326 | .326 | .326 | .326 |
| | $\theta_u^\star$ | | | | .795 | .795 | .795 | .795 | .795 | .720 | .717 | .634 | .515 | .515 | .515 |
| $\text{SD}_t^+(\cdot\|00)$ | $\theta_l^\star$ | n/a | n/a | n/a | .000 | .000 | .000 | .111 | .135 | .410 | .410 | .410 | .410 | .410 | .410 |
| | $\theta_u^\star$ | | | | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | .906 | .902 | .797 | .647 | .647 | .647 |
| $\text{SD}_t^+(\cdot\|000)$ | $\theta_l^\star$ | n/a | n/a | n/a | .000 | .000 | .000 | .000 | .046 | .347 | .393 | .393 | .393 | .393 | .393 |
| | $\theta_u^\star$ | | | | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | .926 | .926 | .926 | .684 | .684 | .684 |
| $\text{SD}_t^-$ | $\theta_l^\star$ | n/a | n/a | n/a | .000 | .000 | .000 | .000 | .000 | .000 | .000 | .000 | .000 | .000 | .000 |
| | $\theta_u^\star$ | | | | .126 | | .099 | .099 | .099 | | | | | | |

Notes: (i) All identified sets are known to be sharp except for those constructed under TIV, i.e. C1, C2 and (10); (ii) Single numbers indicate point identified parameters; (iii) MIV refers to (14); (iv) All parameters are reported for $t=3$.

Table 2L: Lengths of Identified Sets

| | PDBR | CFHN | | DPO | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | (P) | (C1) | (C2) | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) | (11) |
| MTR | n/a | | ✓ | | ✓ | | | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| ST $m = 0$ | n/a | n/a | n/a | | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| ST $m = 1$ | n/a | n/a | n/a | | | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| ST $m = 2$ | n/a | n/a | n/a | | | | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| DSC | n/a | n/a | n/a | | | | | | | ✓ | ✓ | ✓ | ✓ | ✓ |
| MIV | n/a | n/a | n/a | | | | | | | | ✓ | ✓ | ✓ | |
| MTS | n/a | n/a | n/a | | | | | | | | | ✓ | ✓ | ✓ |
| TIV | n/a | ✓ | ✓ | | | | | | | | | | ✓ | |
| $\text{ATE}_t$ | .000 | .618 | .587 | 1.00 | .874 | .906 | .900 | .862 | .653 | .648 | .546 | .268 | .268 | .268 |
| $\text{SD}_t^+$ | n/a | n/a | .587 | .874 | .874 | .853 | .821 | .795 | .653 | .648 | .546 | .268 | .268 | .268 |
| $\text{SD}_t^+(\cdot\|0)$ | n/a | n/a | n/a | .795 | .795 | .795 | .707 | .688 | .394 | .391 | .308 | .188 | .188 | .188 |
| $\text{SD}_t^+(\cdot\|00)$ | n/a | n/a | n/a | 1.00 | 1.00 | 1.00 | .889 | .865 | .496 | .491 | .387 | .237 | .237 | .237 |
| $\text{SD}_t^+(\cdot\|000)$ | n/a | n/a | n/a | 1.00 | 1.00 | 1.00 | 1.00 | .954 | .579 | .533 | .533 | .291 | .291 | .291 |
| $\text{SD}_t^-$ | n/a | n/a | n/a | .126 | .000 | .099 | .099 | .099 | .000 | .000 | .000 | .000 | .000 | .000 |

This table provides a quick comparison of the lengths of the identified sets in Table 2S. (It contains strictly less information than that table.) The same notes for Table 2S apply here.

Table 3: Data Generating Process for the Monte Carlo Study

| $Y = y$ | $\mathbb{P}[Y = y]$ | $Y = y$ | $\mathbb{P}[Y = y]$ |
|---|---|---|---|
| 11111 | .5354 | 00000 | .1481 |
| 01111 | .0329 | 10000 | .0300 |
| 11000 | .0277 | 11110 | .0259 |
| 00111 | .0236 | 11100 | .0207 |
| 00001 | .0207 | 10111 | .0190 |
| 11101 | .0173 | 00011 | .0173 |
| 10001 | .0121 | 11011 | .0110 |
| 11001 | .0104 | 00010 | .0098 |
| 01000 | .0098 | 10011 | .0086 |
| 00100 | .0086 | 01100 | .0046 |
| 00110 | .0035 | 01110 | .0029 |

The probabilities are determined from the empirical distribution of the PSID data (with $T = 4$) after removing all women with more than 2 transitions. This leaves $n = 1{,}735$ cross-sectional observations.

Table 4: Monte Carlo Simulation Results for Specification 1

| | | $\boldsymbol{\theta_l}$ | $\boldsymbol{\theta_u}$ | | rejection probability of $H_0 : t \in \Theta^\star$ for $t = \ldots$ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | true | .080 | .868 | | .000 | .020 | .040 | .060 | .080 | .868 | .875 | .880 | .890 | .900 |
| $n = 1735$ | mean | .086 | .863 | **MR:** | .510 | .140 | .036 | 0 | 0 | 0 | 0 | 0 | .092 | .614 |
| | std | .025 | .008 | **SS:** | .226 | .082 | .024 | 0 | 0 | 0 | .016 | .082 | .572 | .960 |
| | 1%/99% | .038 | .882 | **MQ:** | .510 | .140 | .040 | 0 | 0 | 0 | .016 | .082 | .572 | .960 |
| | min/max | .021 | .884 | **QM:** | .826 | .504 | .252 | .096 | .030 | .014 | .082 | .278 | .790 | .990 |
| $n = 3470$ | mean | .082 | .866 | **MR:** | .932 | .506 | .130 | .008 | 0 | 0 | .006 | .022 | .512 | .984 |
| | std | .019 | .006 | **SS:** | .900 | .478 | .132 | .028 | 0 | .004 | .042 | .196 | .882 | 1 |
| | 1%/99% | .039 | .880 | **MQ:** | .934 | .526 | .140 | .028 | 0 | .004 | .042 | .196 | .882 | 1 |
| | min/max | .025 | .883 | **QM:** | .996 | .862 | .542 | .194 | .036 | .020 | .156 | .482 | .960 | 1 |

Specification 1 imposes MTR and ST with $m = 2$. The parameter of interest is $\text{SD}_t^+$. The time horizon is $T = 4$ and observations with more than 2 transitions are discarded.

Other notes: (i) All tests have nominal level $\alpha = .05$; (ii) The rows %1/99% and min/max give the .01 quantile (and min) across simulations of the direct estimates of $\theta_l^\star$ and the .99 quantile (and max) of the direct estimates of $\theta_u^\star$; (iii) The boxed values of $\theta$ are elements of the identified set; (iv) Tuning parameters are set at $B = 500, \kappa_n = \sqrt{\log(n)}$ and $b_n = n^{2/3}$.

Table 5: Monte Carlo Simulation Results for Specification 2

| | | $\boldsymbol{\theta_l}$ | $\boldsymbol{\theta_u}$ | | rejection probability of $H_0 : t \in \Theta^\star$ for $t = \ldots$ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | true | .080 | .606 | | .000 | .020 | .040 | .060 | .080 | .606 | .630 | .650 | .670 | .690 |
| $n = 1735$ | mean | .086 | .596 | **MR:** | .516 | .144 | .038 | 0 | 0 | 0 | .004 | .015 | .088 | .269 |
| | std | .025 | .034 | **SS:** | .209 | .088 | .019 | .002 | 0 | 0 | 0 | .004 | .017 | .058 |
| | 1%/99% | .038 | .673 | **MQ:** | .516 | .144 | .040 | .002 | 0 | 0 | .004 | .015 | .088 | .269 |
| | min/max | .021 | .687 | **QM:** | .833 | .514 | .253 | .092 | .029 | .008 | .035 | .121 | .296 | .597 |
| $n = 3470$ | mean | .082 | .604 | **MR:** | .932 | .506 | .130 | .008 | 0 | .002 | .030 | .166 | .430 | .826 |
| | std | .019 | .024 | **SS:** | .888 | .478 | .136 | .022 | 0 | 0 | .002 | .014 | .102 | .332 |
| | 1%/99% | .039 | .654 | **MQ:** | .932 | .520 | .150 | .022 | 0 | .002 | .030 | .166 | .430 | .826 |
| | min/max | .025 | .668 | **QM:** | .996 | .858 | .532 | .200 | .036 | .028 | .192 | .454 | .788 | .964 |

Specification 2 imposes MTR, ST with $m = 2$ and MTS. The parameter of interest is $\text{SD}_t^+$. The time horizon is $T = 4$ and observations with more than 2 transitions are discarded.

Other notes: (i) All tests have nominal level $\alpha = .05$; (ii) The rows %1/99% and min/max give the .01 quantile (and min) across simulations of the direct estimates of $\theta_l^\star$ and the .99 quantile (and max) of the direct estimates of $\theta_u^\star$; (iii) The boxed values of $\theta$ are elements of the identified set; (iv) Tuning parameters are set at $B = 500, \kappa_n = \sqrt{\log(n)}$ and $b_n = n^{2/3}$.

## A Extension to Discrete Outcomes

The DPO model extends readily to the case where $Y_t$ assumes values in $\{0, 1, \ldots, J\}$ with $J > 1$ and hence $Y$ 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. In this more general case, there are $J + 1$ counterfactual outcomes $\{U_t(y)\}_{y=0}^{J}$ for each $t \geq 1$. The observed outcome in period $t$ is determined as

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

The structure $P$ is a probability mass function for $(Y_0, \{U_t(0), \ldots, U_t(J)\}_{t=1}^{T})$ and the characterization of the identified set remains unchanged. Parameters and auxiliary identifying assumptions that are appropriate for the $J = 1$ case may or may not be appropriate for the $J > 1$ case and vice-versa, but a separate analysis is beyond the scope of this paper.

## B Extension to Higher Order State Dependence

The discussion in the main text is premised on the assumption 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 setting is consistent with much of the empirical and theoretical literature on state (vs. duration) dependence. In this section I outline how one would extend the model to allow for state dependence of higher orders.

When $Y_t$ is binary, this generalization to state dependence of length $K \geq 1$ is accomplished by introducing $2^K$ counterfactual outcomes $\{U_t(y)\}_{y \in \{0,1\}^K}$ for each period $t \geq K$. The recursive relationship (6) is replaced by

$$Y_t = \sum_{y \in \{0,1\}^K} U_t(y) \mathbb{1}[(Y_t, Y_{t-1}, \ldots, Y_{t-K+1}) = y] \quad \text{for } t \geq K, \tag{22}$$

with the joint determination of periods $t = 0$ up to $t = K - 1$ not being modeled explicitly. For example, with $K = 2$, (22) would become

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

so that for each $t$ there are four potential outcomes corresponding to the four potential two-period histories immediately prior to period $t$.

The structure $P$ is a probability mass function for the random vector

$$\left(Y_0, Y_1, \ldots, Y_{K-1}, \left\{U_t(y) : y \in \{0,1\}^K\right\}_{t=K}^T\right).$$

The identified set $\mathcal{P}^\star$ can be characterized through essentially the same argument as for the first order case. Parameters and auxiliary identifying assumptions would need to be reconsidered for the higher order case.

## C Derivation of Bounds Using Only The Empirical Evidence

Here I justify the claim that (11) and (12) are sharp bounds for $\mathrm{SD}_t^+$ and $\mathrm{SD}_t^-$. Observe that if $P \in \mathcal{P}^\star$ then

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

where the second equality follows because under (6), $[Y_{t-1}=0, U_t(0)=0]$ if and only if $[Y_{t-1}=0, Y_t=0]$ and $[Y_{t-1}=1, U_t(1)=1]$ if and only if $[Y_{t-1}=1, Y_t=1]$. The only restrictions implied on the second two terms are

$$\begin{aligned}
0 \geq -\mathbb{P}_P[Y_{t-1}=0, Y_t=0, U_t(1)=0] \geq -\mathbb{P}_P[Y_{t-1}=0, Y_t=0] \\
0 \geq -\mathbb{P}_P[Y_{t-1}=1, U_t(0)=1, Y_t=1] \geq -\mathbb{P}_P[Y_{t-1}=1, Y_t=1],
\end{aligned} \tag{23}$$

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 (23), and one obtaining both of the lower bounds. The upper and lower bounds in (11) now follow from those in (23). The bounds in (12) follow from an analogous argument using the decomposition

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

## D 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 was shown for $\text{SD}_t^+$ in (10), but not for any subsequent parameters or assumptions. This section contains some additional discussion. For simplicity, I assume throughout that $X$ is degenerate, but it is straightforward to adjust the conditions to allow for $X$ to be random by simply conditioning and then averaging over all realizations of $X$.

To see that $\text{SD}_t^+(P|0)$ is linear, write it as

$$\text{SD}_t^+(P|0) = \frac{\mathbb{P}_P[U_t(0) = 0, U_t(1) = 1, Y_t = 0]}{\mathbb{P}[Y_t = 0]} = \frac{\sum_{u \in \mathcal{U}_t^+(0)} P[u]}{\mathbb{P}[Y_t = 0]},$$

where $\mathcal{U}_t^+(0)$ is the set of $u \in \mathcal{U}$ such that $u_t(0) = 0, u_t(1) = 1$, and $Y_t = 0$ when computed through the recursive relationship (6) with $Y_0 = u_0$, $U_t(0) = u_t(0)$ and $U_t(1) = u_t(1)$. Similar equations follow for $\text{SD}_t^+(P|00), \text{SD}_t^+(P|000)$ and any other outcome-conditioned parameter. The division by an observed probability in these expressions does not introduce a nonlinearity, because any $P \in \mathcal{P}^\star$ must satisfy $\mathbb{P}_P[Y_t = 0] = \mathbb{P}[Y_t = 0]$ in order to be observationally equivalent.

Assumption MTR is linear because it can be written as $\mathbb{P}_P[U_t(0) = 1, U_t(1) = 0] = 0$ for all $t \geq 1$. Hence, let $\mathcal{U}_t^{\text{MTR}}$ denote the set of all $u \in \mathcal{U}$ such that $u_t(1) = 0$ and $u_t(0) = 1$, and then write MTR as

$$\sum_{u \in \mathcal{U}_t^{\text{MTR}}} P[u] = 0, \tag{24}$$

for all $t \geq 1$. In terms of the $\rho$ function, this equality constraint can be imposed with two inequalities. Assumptions ST, MIV and TIV can be imposed similarly by summing over the appropriate sub-collections of $\mathcal{U}$. Assumption MTS can be imposed using a construction similar to that for $\text{SD}_t^+(P|00)$.

Now, consider Assumption DSC. In general, this is a nonlinear restriction, but if ST holds so that the distribution of $U_t(d)$ does not depend on $t$, then $\text{Corr}(U_t(d), U_{t+s}(d))$ is decreasing in $|s|$ if and only if $\text{Cov}(U_t(d), U_{t+s}(d))$ is decreasing in $|s|$. Furthermore, under ST, the latter is true if and only if $\mathbb{E}[U_t(d)U_{t+s}(d)]$, i.e. $\mathbb{P}[U_t(d) = 1, U_{t+s}(d) = 1]$, is decreasing in $|s|$. It is straightforward to show that $\mathbb{P}[U_t(d) = 1, U_{t+s}(d) = 1]$ is a linear function of $P$ using arguments like in (24).

## E Proof of Proposition 1

Suppose that $V_t|A, Y_0$ forms a first-order Markov chain and that $\mathbb{P}[V_{t+1} \leq v_{t+1}|V_t = v_t, A, Y_0]$ is decreasing in $v_t$ for all $v_{t+1}$. By Theorem 3.1 of Fang et al. (1994), $(V_t, V_{t+s})$ is decreasing in the concordance ordering as a function of $|s|$, conditional on $A, Y_0$. That is, $\mathbb{P}[V_t \geq v_1, V_{t+s} \geq v_2|A, Y_0]$ is decreasing in $|s|$ for any $v_1, v_2$. Hence, for any integers $s, s'$ with $|s| < |s'|$

$$\begin{aligned}
&\mathbb{P}[U_t(y) = 1, U_{t+s}(y) = 1] \\
&\quad = \mathbb{E}\left[\mathbb{P}[U_t(y) = 1, U_{t+s}(y) = 1|A, Y_0]\right] \\
&\quad = \mathbb{E}\left[\mathbb{P}[\gamma y + \lambda Y_0 + A + V_t \geq 0, \gamma y + \lambda Y_0 + A + V_{t+s} \geq 0|A, Y_0]\right] \\
&\quad \geq \mathbb{E}\left[\mathbb{P}[\gamma y + \lambda Y_0 + A + V_t \geq 0, \gamma y + \lambda Y_0 + A + V_{t+s'} \geq 0|A, Y_0]\right] \\
&\quad = \mathbb{E}\left[\mathbb{P}[U_t(y) = 1, U_{t+s'}(y) = 1|A, Y_0]\right] = \mathbb{P}[U_t(y) = 1, U_{t+s'}(y) = 1],
\end{aligned}$$

which implies DSC, given ST.

## F Proof of Proposition 2

For $y \in \{0, 1\}$ and $t = 0, \ldots, T-1$ let $\xi_t^y$ denote an indicator variable for the event that $Y_t = y$ but $Y_s \neq y$ for all $0 \leq s \leq t-1$. Also, let $\bar{\xi}^y$ denote an indicator variable for the event that $Y_t \neq y$ for all $t \leq T-1$. Then $\sum_{t=0}^{T-1} \xi_t^y + \bar{\xi}^y = 1$, since the events these indicators indicate for are disjoint and exhaustive. Hence

$$\begin{aligned}
\mathbb{E}_P[U_1(y)] &= \mathbb{E}_P\left[\mathbb{E}_P[U_1(y)|Y_0, A]\right] \\
&= \sum_{t=0}^{T-1} \mathbb{E}_P\left[\xi_t^y \, \mathbb{E}_P[U_1(y)|Y_0, A]\right] + \mathbb{E}_P\left[\bar{\xi}^y \, \mathbb{E}_P[U_1(y)|Y_0, A]\right].
\end{aligned}$$

For each $t = 0, \ldots, T-1$, TIV implies that

$$\begin{aligned}
\mathbb{E}_P\left[\xi_t^y \, \mathbb{E}_P[U_1(y)|Y_0, A]\right] &= \mathbb{E}_P\left[\xi_t^y \, \mathbb{E}_P[U_{t+1}(y)|Y_t, \ldots, Y_0, A]\right] \\
&= \mathbb{E}_P\left[\mathbb{E}_P[\xi_t^y U_{t+1}(y)|Y_t, \ldots, Y_0, A]\right] = \mathbb{E}_P\left[\xi_t^y U_{t+1}(y)\right],
\end{aligned}$$

where the second equality follows because $\xi_t^y$ is a function of $Y_t, \ldots, Y_0$. A similar argument shows that

$$\mathbb{E}_P\left[\bar{\xi}^y \, \mathbb{E}_P[U_1(y)|Y_0, A]\right] = \mathbb{E}_P\left[\bar{\xi}^y U_T(y)\right].$$

Since $\xi_t^y = 1$ implies $Y_{t+1} = U_{t+1}(Y_t) = U_{t+1}(y)$, one has for any $P \in \mathcal{P}^\star$,

$$\begin{aligned}\mathbb{E}_P[U_1(y)] &= \sum_{t=0}^{T-1} \mathbb{E}_P\left[\xi_t^y Y_{t+1}\right] + \mathbb{E}_P\left[\bar{\xi}^y U_T(y)\right] \\ &= \sum_{t=0}^{T-1} \mathbb{P}[Y_{t+1} = 1, Y_t = y, Y_s \neq y \; \forall s \leq t-1] \\ &\qquad + \mathbb{E}_P\left[U_T(y) | Y_s \neq y \; \forall s \leq T-1\right] \mathbb{P}[Y_s \neq y \; \forall s \leq T-1].\end{aligned}$$

Setting $\mathbb{E}_P\left[U_T(y) | Y_s \neq y \; \forall s \leq T-1\right]$ to 0 or 1 delivers the asserted bounds for $t = 1$. Observing that TIV implies

$$\mathbb{E}_P[U_t(y)] = \mathbb{E}_P\left[\mathbb{E}_P[U_t(y)|Y_{t-1}, \ldots, Y_0, A]\right] = \mathbb{E}_P\left[\mathbb{E}_P[U_1(y)|Y_0, A]\right] = \mathbb{E}_P[U_1(y)],$$

shows that the bounds hold for all $t$.

## References


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.

Andersen, E. B. (1970): "Asymptotic Properties of Conditional Maximum-Likelihood Estimators," *Journal of the Royal Statistical Society. Series B (Methodological)*, 32, 283–301.

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.

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

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., I. Canay, and X. Shi (2014): "Inference for Functions of Partially Identified Parameters in Moment Inequality Models," *cemmap working paper 22/14*.

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

Byrd, R. H., J. Nocedal, and R. A. Waltz (2006): "KNITRO: An integrated package for nonlinear optimization," in *Large-scale nonlinear optimization*, Springer, 35–59.

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

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.

——— (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*.

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.

Chesher, A. (2010): "Instrumental Variable Models for Discrete Outcomes," *Econometrica*, 78, 575–601.

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

Ciliberto, F. and E. Tamer (2009): "Market Structure and Multiple Equilibria in Airline Markets," *Econometrica*, 77, 1791–1828.

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.

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.

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.

Eckstein, Z. and O. Lifshitz (2011): "Dynamic Female Labor Supply," *Econometrica*, 79, 1675–1726.

Eckstein, Z. and K. I. Wolpin (1989): "Dynamic Labour Force Participation of Married Women and Endogenous Work Experience," *The Review of Economic Studies*, 56, 375–390.

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.

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.

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

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

Freyberger, J. and J. Horowitz (2013): "Identification and Shape Restrictions in Nonparametric Instrumental Variables Estimation," *cemmap working paper CWP31/13*.

Gandhi, A., Z. Lu, and X. Shi (2013): "Estimating Demand for Differentiated Products with Error in Market Shares," *Working paper*.

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

Gronau, R. (1974): "Wage Comparisons–A Selectivity Bias," *Journal of Political Economy*, 82, 1119–1143.

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.

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

Heckman, J. (1974): "Shadow Prices, Market Wages, and Labor Supply," *Econometrica*, 42, 679–694.

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): "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 T. E. MaCurdy (1980): "A Life Cycle Model of Female Labour Supply," *The Review of Economic Studies*, 47, 47–74.

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 S. Urzua (2010): "Comparing IV with structural models: What simple IV can and cannot identify," *Journal of Econometrics*, 156, 27–37.

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. 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.

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

Hurwicz, L. (1950): "Generalization of the Concept of Identification," in *Statistical Inference in Dynamic Economic Models*, ed. by T. Koopmans, no. 10 in Cowles Commission Monographs.

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

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. and R. M. Sauer (2009): "Classification Error in Dynamic Discrete Choice Models: Implications for Female Labor Supply Behavior," *Econometrica*, 77, 975–991.

Khan, S., M. Ponomareva, and E. Tamer (2011): "Sharpness in randomly censored linear models," *Economics Letters*, 113, 23–25.

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. (2015): "Bounding Average Treatment Effects Using Linear Programming," *Working paper*.

Lee, D. S. (2008): "Randomized experiments from non-random selection in U.S. House elections," *Journal of Econometrics*, 142, 675–697.

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

Manski, C. F. (1975): "Maximum score estimation of the stochastic utility model of choice," *Journal of Econometrics*, 3, 205–228.

——— (1989): "Anatomy of the Selection Problem," *The Journal of Human Resources*, 24, 343–360.

——— (1996): "Learning about Treatment Effects from Experiments with Random Assignment of Treatments," *The Journal of Human Resources*, 31, 709–733.

——— (1997a): "Monotone Treatment Response," *Econometrica*, 65, 1311–1334.

——— (1997b): "The Mixing Problem in Programme Evaluation," *The Review of Economic Studies*, 64, 537–553.

——— (2003): *Partial identification of probability distributions*, Springer.

——— (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 incomeleisure 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.

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

Nelsen, R. (2006): *An introduction to copulas*, New York: Springer.

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

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

Pakes, A. and J. Porter (2014): "Moment Inequalities for Multinomial Choice with Fixed Effects," *Working paper*.

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

Rasch, G. (1961): "On general laws and the meaning of measurement in psychology," in *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, University of California Press Berkeley, CA, vol. 4, 321–333.

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.

Shiu, J.-L. and Y. Hu (2013): "Identification and estimation of nonlinear dynamic panel data models with unobserved covariates," *Journal of Econometrics*, 175, 116–131.

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.