Title
Testing identifying assumptions in nonseparable panel data models

Journal
Journal of Econometrics, 197(2)

ISSN
0304-4076

Author
Ghanem, Dalia

Publication Date
2017-04-01

DOI
10.1016/j.jeconom.2016.11.005

Peer reviewed

Testing Identifying Assumptions in Nonseparable Panel Data Models

Dalia Ghanem*

October 26, 2016

Abstract

Recent work on nonparametric identification of average partial effects (APEs) from panel data require restrictions on individual or time heterogeneity. Identifying assumptions under the "generalized first-differencing" category, such as time homogeneity (Chernozhukov, Fernandez-Val, Hahn, and Newey, 2013), have testable equality restrictions on the distribution of the outcome variable. This paper proposes specification tests based on these restrictions. The bootstrap critical values for the resulting Kolmogorov-Smirnov and Cramer-von-Mises statistics are shown to be asymptotically valid and deliver good finite-sample properties in Monte Carlo simulations. An empirical application illustrates the merits of testing nonparametric identification from an empiricist's perspective.

JEL: C1, C14, C21, C23, C25

Keywords: panel data, nonparametric identification, specification testing, discrete regressors, bootstrap adjustment, Kolmogorov-Smirnov statistic, Cramer-von-Mises statistic

*Department of Agricultural and Resource Economics, University of California, Davis, One Shields Ave, Davis CA 95616, Phone: 530-752-2414, Fax: 530-752-5614, e-mail: dghanem@ucdavis.edu

1 Introduction

In many empirical settings in economics, panel data are used to identify the effect of a regressor on an outcome of interest. To ensure that a panel data set can identify such an effect, empirical economists typically utilize a number of robustness checks or test the parallel-trends assumption, when there are more than two time periods available.[1] This paper proposes alternative specification tests that build on recent developments in the theoretical literature on nonparametric identification of average partial effects (APEs) from panel data. As a key advantage, these new tests do not rely on parametric assumptions which are likely to be misspecified in practice. Furthermore, the tests proposed here can be applied when only two time periods are available, since they exploit restrictions on the entire distribution of the outcome variable as opposed to restrictions on its mean only, as in the case of the parallel-trends assumption.

Recent work on nonparametric identification of APEs from fixed-T panel data extend fixed effects and correlated random effects identification strategies, originally introduced in the linear model (Chamberlain, 1984; Mundlak, 1978), to fully nonseparable models. The identifying assumptions in this setting may be viewed as restrictions on the following structural relationship,

$$Y_{it} = \xi_t(X_{it}, \mathcal{A}_i, \mathcal{U}_{it}), \text{ for } i = 1, \ldots, n \text{ and } t = 1, \ldots, T. \tag{1}$$

Y_{it} is the outcome variable of interest, X_{it} is a $d_x \times 1$ regressor vector, which is assumed to have finite support. Hence, the testing procedures proposed in this paper can only allow discrete regressors to enter the structural function in a fully nonseparable way as in the above model.[2] $\mathcal{A}_i$ and $\mathcal{U}_{it}$ are individual-specific time-invariant and time-varying unobservables, respectively. The model is static, hence X_{it} does not include lagged dependent variables or other variables that can introduce feedback mechanisms. Equation (1) reflects the threats to identification that researchers are typically concerned with in the panel data context. The

[1]The test of the parallel-trends assumption is most commonly used when the regressor is binary as in the difference-in-difference framework.

[2]Section A in the supplementary appendix allows for a linear index of additively separable continuous regressors.

time-invariant and time-varying unobservables may vary with the regressors and confound the effect of interest. Furthermore, the relationship between the outcome variables, regressors and unobservables, i.e., the structural function ξ_t, may change over time. Without further restrictions, we cannot identify APEs from panel data.

Fixed effects strategies in nonseparable panel data models impose time homogeneity assumptions, which restrict changes in the structural function and the distribution of the idiosyncratic shocks across time (Chernozhukov, Fernandez-Val, Hahn, and Newey, 2013; Hoderlein and White, 2012). Correlated random effects strategies restrict individual heterogeneity (Altonji and Matzkin, 2005; Bester and Hansen, 2009). Fortunately, the identifying assumptions in both cases imply testable equality restrictions on the conditional distribution of the outcome variable. The contribution of this paper is to develop specification tests based on these restrictions for time homogeneity in the presence of parallel trends as well as correlated random effects. The testing problem extends the classical two-sample problem in statistics to the panel setting, where the two samples are dependent and the data are possibly demeaned. We hence propose bootstrap-adjusted critical values for Kolmogorov-Smirnov (KS) and Cramer-von-Mises (CM) statistics and show that they are asymptotically valid.

From a practical standpoint, the choice over which identifying assumption to test emerges naturally from the empirical context. The test for time homogeneity is suitable for observational panel data settings, where empirical researchers prefer fixed effects approaches allowing for arbitrary individual heterogeneity. The intuition behind the testable restriction is that subpopulations that do not experience changes in X over time (stayers) should have the same distribution of the outcome variable after accounting for the parallel trend. Observational panel data are the most widely used panel data in economics. However, the recent growth of field experiments, where both baseline (pre-treatment) and follow-up (post-treatment) outcome variables are collected, gives rise to a new type of panel data, which we refer to as "experimental panel data." For this type of panel data, identification of the APE, which is more commonly referred to as the treatment on the treated (TOT), is achieved through the conditional random effects assumption. This assumption falls under the correlated random effects category and relates to unconfoundedness. Treatment is randomly assigned in the second time period. Therefore, we expect that the treatment and control group have the same baseline distribution for the outcome variable if no implementation issues, such as attrition or other selection problems, interfere with the randomization. Hence, this test can be tailored to test identification in the presence of selection problems.

One of the main advantages of the specification tests proposed here is the interpretation of their rejection. Unlike over-identification tests that combine several assumptions, the specification tests here are based on implications of a particular identifying assumption. As a result, their rejection is clear evidence against the identifying assumption in question. It is important to note however that a rejection does not mean that the APE is not identified, but rather that it is not identified by means of the chosen identifying assumption. Furthermore, due to the nonparametric and nonseparable nature of the structural function we consider, there are no additional functional-form assumptions that may be misspecified and hence drive a rejection of the test, which is a problem with over-identification tests that rely on parametric assumptions. This issue will be illustrated in an empirical example revisiting Angrist and Newey (1991).

Another advantage of the specification tests proposed here is that they do not require over-identifying restrictions on the object of interest, the APE. One practical implication is that some of the specification tests, such as time homogeneity and conditional random effects, are applicable even when $T = 2$. The commonly used test of the parallel-trends assumption requires that $T > 2$. This is because the specification tests proposed here test distributional assumptions, whereas the parallel trends assumption only exploits mean restrictions.

Related Literature. This paper builds on the recent work on fixed-T nonparametric identification of APEs, which we categorize as "generalized first-differencing." Under this category, the APE is identified for a subpopulation using average changes of the outcome variable across time or subpopulations that coincide with a change in the variable of interest. This generalization allows us to put correlated random effects (Altonji and Matzkin, 2005; Bester and Hansen, 2009) and time homogeneity (Chernozhukov, Fernandez-Val, Hahn, and Newey, 2013; Hoderlein and White, 2012) under the same umbrella. Point-identification in these papers is achieved only for a subpopulation. Chernozhukov, Fernandez-Val, Hahn, and

Newey (2013) are not solely interested in the point-identified subpopulation, but are mainly concerned with set identification and estimation building on Honore and Tamer (2006) and Chernozhukov, Hong, and Tamer (2007). Another strand in the literature follows the classical identification approach, which seeks to identify all structural objects, i.e., the structural function and the conditional distribution of unobservables. This strand includes Altonji and Matzkin (2005), Evdokimov (2010) and Evdokimov (2011).

The study of identification in panel data originated in the linear model with the seminal work of Mundlak (1978), Chamberlain (1982), and Chamberlain (1984). The latter together with more recent surveys such as Arellano and Honore (2001) and Arellano and Bonhomme (2011) discuss the role of separability and time homogeneity in the linear as well as nonlinear models. Variants of time homogeneity are also assumed in other work on nonlinear panel data models, such as in Manski (1987) and Honore and Kyriazidou (2000b) for binary choice models, Honore (1992), Honore (1993), Kyriazidou (1997), and Honore and Kyriazidou (2000a) for Tobit models.[3]

It is important to relate the differencing approach here to other identification approaches in the literature. For random coefficient models, Graham and Powell (2012) use a differencing approach that is similar in spirit to what we use here. Magnac (2004) introduces the concept of quasi-differencing as the presence of a sufficient statistic for the individual effect, such as conditional logit (Chamberlain, 1984, 2010). For semiparametric binary choice models, Honore and Kyriazidou (2000b) use the intuition of quasi-differencing to nonparametrically identify the common parameter. In the likelihood setting, Bonhomme (2012) proposes a systematic approach to finding moment restrictions that only depend on the common parameter vector, and refers to it as a "functional differencing" approach.

In the nonlinear difference-in-difference setup with repeated cross-sections, as in Athey and Imbens (2006) for binary treatments, and d'Haultfoeuille, Hoderlein, and Sasaki (2013) for continuous treatments, variants of time homogeneity are imposed while allowing the structural relationship to be time-varying. These identification strategies do not fall under the

[3]Note for Tobit models, exchangeability of the error terms across time is assumed, which is an implication of time homogeneity as assumed in Chernozhukov, Fernandez-Val, Hahn, and Newey (2013).

generalized first-differencing approach.[4] Their identification approach relies on monotonicity assumptions. Athey and Imbens (2006) exploit the monotonicity of the structural function in a scalar unobservable, together with time homogeneity, to identify the distribution of the counterfactual from which they can derive the APE. In d'Haultfoeuille, Hoderlein, and Sasaki (2013), the time-varying structural function is monotonic in a time-invariant base function of the regressors and unobservables. Together with a variant of time homogeneity and a restriction on the variation in the distribution of the regressors across time, d'Haultfoeuille, Hoderlein, and Sasaki (2013) point-identify several types of average and quantile effects.

The motivation behind the specification tests proposed here is most closely related to recent work on nonparametric specification testing. Su, Hoderlein, and White (2013) and Hoderlein, Su, and White (2013) propose nonparametric tests of monotonicity in a scalar unobservable in the panel and cross-sectional setup, respectively. Hoderlein and Mammen (2009), Lu and White (2014) and Su, Tu, and Ullah (2015) propose nonparametric tests of separability. This paper also builds on work using re-sampling methods to obtain critical values for KS statistics such as Andrews (1997) and Abadie (2002).

Outline of the Paper. The rest of this paper is organized as follows. The next section reviews the identifying assumptions considered here and their testable restrictions. Section 3 proposes tests of the identifying assumptions that are shown to be asymptotically valid and presents a simulation study to examine their performance in finite samples. Finally, Section 4 includes the empirical illustration.

2 Generalized First-Differencing: Identification and Testable Restrictions

In this section, we review the identifying assumptions that fall under the umbrella of generalized first-differencing and present testable restrictions. We start from the DGP in (1) and

[4] This is not surprising, since in repeated cross-sections, we do not observe the same individuals across time. Hence, we cannot identify the APE from average changes across time as in the generalized first-differencing approach, if the underlying structural function is fully nonseparable.

formally state the main assumptions. Let $X_i = (X_{i1}, \ldots, X_{iT})$. For a set $\mathbb{S}$, $|\mathbb{S}|$ denotes its cardinality and $\mathbb{S}^T = \times_{t=1}^{T}\mathbb{S}$, e.g., $\mathbb{S}^2 = \mathbb{S} \times \mathbb{S}$.

Assumption 2.1 *(General DGP)*

(i) $Y_{it} = \xi_t(X_{it}, \mathcal{A}_i, \mathcal{U}_{it})$, *where* $Y_{it} \in \mathbb{Y} \subseteq \mathbb{R}$, $X_{it} \in \mathbb{X}$, $\mathcal{A}_i \in \mathbb{A} \subseteq \mathbb{R}^{d_a}$, $\mathcal{U}_{it} \in \mathbb{U} \subseteq \mathbb{R}^{d_u}$, *for* $t = 1, 2, ..., T$,

(ii) $\mathbb{X}$ *is finite, where* $|\mathbb{X}| = K$,

(iii) $E[Y_{it}] < \infty$ *for all* $t = 1, 2, .., T$,

(iv) $P(X_i = \underline{x}) > 0$ *for all* $\underline{x} \in \mathbb{X}^T$.

The main content of the above assumption is the finite support of X_{it} in (ii). (i) may be thought of as a 'correct specification' assumption. However, the choice of variables to include in X_{it} only becomes restrictive when an identifying assumption is imposed on (i). As for (iii) and (iv), they are regularity conditions that ensure that the APE exists for all elements in the support of X_i, which simplifies our analysis. As noted above, we assume that $\{Y_{it}, X_{it}\}$ are observable and $\{\mathcal{A}_i, \mathcal{U}_{it}\}$ are unobservable, where $\mathcal{A}_i$ and $\mathcal{U}_{it}$ may be any finite-dimensional vectors.

For the purposes of our discussion here, we define a subpopulation by its realization of $X_i = \underline{x}$, where $\underline{x} \in \mathbb{X}^T$.[5] Since $|\mathbb{X}^T|$ is finite, we have finitely many subpopulations. Each subpopulation, $\underline{x} \in \mathbb{X}^T$, is characterized by its distribution of unobservables, i.e. $F_{\mathcal{A}_i,\mathcal{U}_{i1},\mathcal{U}_{i2},\ldots,\mathcal{U}_{iT}|X_i}(.|\underline{x})$. Hence, we can think of individuals in a subpopulation as draws from the same distribution.[6]

Now we define our object of interest using the general DGP. For a fixed-T panel, the APE is only point-identified for a subpopulation as established previously in Chernozhukov, Fernandez-Val, Hahn, and Newey (2013) and Hoderlein and White (2012). Hence, our object

[5]This is in line with common approaches in the literature to divide up the populations into groups based on the realizations of the regressors or treatment variables, such as movers and stayers as introduced in Chamberlain (1982), and the treatment and control groups in the treatment effects literature.

[6]For fixed effects strategies, one can allow $F_{\mathcal{A}_i,\mathcal{U}_{i1},\mathcal{U}_{i2},\ldots,\mathcal{U}_{iT}|X_i}(.|\underline{x}) = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} F^i_{\mathcal{A}_i,\mathcal{U}_{i1},\mathcal{U}_{i2},\ldots,\mathcal{U}_{iT}|X_i}(.|\underline{x})$, i.e., each individual in a subpopulation may be a draw from a different distribution. The intuition here is that we are using within-group variation, individual heterogeneity could be even more general. However, for correlated random effects strategies, this will interfere with identification.

of interest here is the APE of a discrete regressor X on Y for a subpopulation $\underline{x}$. Specifically, our object of interest is the APE of changing X_{it} from x to x', $x \neq x'$, for subpopulation $X_i = \underline{x}$. We use the counterfactual notation, $Y_{it}^{x} = \xi_t(x, \mathcal{A}_i, \mathcal{U}_{it})$.

$$\begin{aligned}\beta_t(x \to x' | X_i = \underline{x}) &= E[Y_{it}^{x'} | X_i = \underline{x}] - E[Y_{it}^{x} | X_i = \underline{x}] \\ &= \int \big\{ \xi_t(x', a, u) - \xi_t(x, a, u) \big\} dF_{\mathcal{A}_i, \mathcal{U}_{it} | X_i}(a, u | \underline{x}). \end{aligned} \tag{2}$$

The above equation expresses the identification problem here. The APE is the difference between the same function ξ_t evaluated at x and x' averaged over the same distribution of unobservables.

The identifying assumptions that fall under the generalized first-differencing category impose restrictions that ensure that the APE is identified by looking at average differences of the outcome variable across time and subpopulations. To simplify the illustration of results, we will focus on the two-period case ($T = 2$), where we have two subpopulations (x, x) and (x, x'), stayers and movers. If $x = 0$ and $x' = 1$, then this would be the classical difference-in-difference setup. Our object of interest is $\beta_2(x \to x' | X_i = (x, x'))$, i.e., it is the APE of moving from x to x' for subpopulation (x, x') in the second time period. In this setup, generalized first-differencing allows for the identification of the APE using $E[Y_{i2} - Y_{i1} | X_i = (x, x)]$ and $E[Y_{i2} - Y_{i1} | X_i = (x, x')]$. The following lemma gives a condition that formally characterizes the generalized first-differencing approach.

Lemma 2.1 ***(Generalized First-Differencing)***
Let Assumption 2.1 hold,

$$\beta_2(x \to x' | X_i = (x, x')) = E[Y_{i2} - Y_{i1} | X_i = (x, x')] - E[Y_{i2} - Y_{i1} | X_i = (x, x)]$$

if and only if

$$\begin{aligned} &\int (\xi_2(x, a, u_2) - \xi_1(x, a, u_1)) \\ &\quad \times \big(dF_{\mathcal{A}_i, \mathcal{U}_{i1}, \mathcal{U}_{i2} | X_i}(a, u_1, u_2 | (x, x')) - dF_{\mathcal{A}_i, \mathcal{U}_{i1}, \mathcal{U}_{i2} | X_i}(a, u_1, u_2 | (x, x)) \big) = 0. \end{aligned} \tag{3}$$

All proofs of Section 2 are given in Appendix A.1. The term on the left-hand side of (3) is the integral of the product of the change in the structural function due to time and the difference in the distribution of unobservables between the two subpopulations. The condition in (3) can be viewed as an orthogonality condition, in the sense that two variables, A and B, are orthogonal if $E[AB] = 0$. Assuming that the density $f_{\mathcal{A}_i,\mathcal{U}_{i1},\mathcal{U}_{i2}|X_i}$ exists and is positive everywhere, we can obtain the following version of the condition,

$$E\left[(Y_{i2}^x - Y_{i1}^x)\frac{\{f_{\mathcal{A}_i,\mathcal{U}_{i1},\mathcal{U}_{i2}|X_i}(\mathcal{A}_i,\mathcal{U}_{i1},\mathcal{U}_{i2}|(x,x')) - f_{\mathcal{A}_i,\mathcal{U}_{i1},\mathcal{U}_{i2}|X_i}(\mathcal{A}_i,\mathcal{U}_{i1},\mathcal{U}_{i2}|(x,x))\}}{f_{\mathcal{A}_i,\mathcal{U}_{i1},\mathcal{U}_{i2}|X_i}(\mathcal{A}_i,\mathcal{U}_{i1},\mathcal{U}_{i2}|(x,x'))}\right| X_i = (x,x')\Bigg] = 0.$$

The above equation shows how (3), while holding x fixed, decomposes the change in the outcome variable into two components: (1) the change in the structural function due to time, (2) the difference in the distribution of unobservables due to the change in X_{i2}. For identification to be achieved here, these two sources of change have to be orthogonal.

The following theorem shows that time homogeneity fulfills the condition (3) and gives the testable restrictions implied by this identification approach. The testable restrictions here are not over-identifying restrictions on the object of interest, the APE.

Theorem 2.1 ***Fixed Effects: Identification & Testable Restrictions*** ***(**$T = 2$***)***

Let Assumption 2.1 hold.

$$\begin{aligned}
&\textit{If} \quad \xi_t(x,a,u) = \xi(x,a,u) + \lambda_t(x), \forall (x,a,u) \in \mathbb{X}\times\mathbb{A}\times\mathbb{U}, \ \textit{and} \\
&\qquad \mathcal{U}_{i1}|X_i,\mathcal{A}_i \overset{d}{=} \mathcal{U}_{i2}|X_i,\mathcal{A}_i, \\
&\textit{then } (i) \int (\xi_2(x,a,u_2) - \xi_1(x,a,u_1)) \\
&\qquad\qquad \left(dF_{\mathcal{A}_i,\mathcal{U}_{i1},\mathcal{U}_{i2}|X_i}(a,u_1,u_2|(x,x')) - dF_{\mathcal{A}_i,\mathcal{U}_{i1},\mathcal{U}_{i2}|X_i}(a,u_1,u_2|(x,x))\right) = 0. \\
&\qquad (ii) F_{Y_{i1}-\lambda_1(x)|X_i}(.|(x,x)) = F_{Y_{i2}-\lambda_2(x)|X_i}(.|(x,x)), \forall x \in \mathbb{X}
\end{aligned}$$

The identity of distribution assumption on the idiosyncratic shocks across time conditional on X_i and $\mathcal{A}_i$ was referred to in Chernozhukov, Fernandez-Val, Hahn, and Newey (2013) as time homogeneity in the case of discrete regressors. Hoderlein and White (2012) also rely on a similar time homogeneity assumption to identify local average structural derivatives.

As pointed out in Chernozhukov, Fernandez-Val, Hoderlein, Holzmann, and Newey (2015), the time homogeneity assumption can be equivalently stated as $\mathcal{E}_{i1}|X_i \stackrel{d}{=} \mathcal{E}_{i2}|X_i$, where $\mathcal{E}_{it} = (\mathcal{A}_i', \mathcal{U}_{it}')'$. The assumption on the structural function ensures that it is stationary in the unobservables. Both Hoderlein and White (2012) and Chernozhukov, Fernandez-Val, Hahn, and Newey (2013) also impose time stationarity assumptions on the structural function. We will refer to $\lambda_t(x)$ as a generalized parallel trend.

The above theorem shows that time homogeneity together with the restriction on the structural function satisfies the generalized first-differencing condition restated in (i) of the above theorem. The identification approach using time homogeneity fits the setup in observational panel data models, where empirical researchers prefer to leave the distribution of unobservable individual heterogeneity unrestricted (fixed effects). The stayer subpopulation (x, x) is used to identify the generalized parallel trend. The difference between the average change of the outcome variable for the mover subpopulation (x, x') and the generalized parallel trend then identifies the APE in question. In essence, time homogeneity together with stationarity of the structural function in unobservables ensures that the distribution of the outcome variable does not change across time due to unobservables. Hence, we can identify the APE from an average change in the outcome variable that coincides with a change in the regressors, after accounting for the parallel trend.

The testable restriction in (ii) of the above theorem follows intuitively. The distribution of the outcome variable for individuals who do not experience changes in their regressors, i.e., the stayer subpopulations, should not change across time after adjusting for the generalized parallel trend.[7] The extension to the case where $T > 2$ is straightforward and is presented in Section 3.1.

The fixed effects approach is particularly appropriate for observational panel data settings. The finite-support assumption on X_{it} can restrict the applicability of the test procedures proposed here. We can allow for a vector of additively separable possibly continuous regressors W_{it}, a $d_w \times 1$ vector, where $Y_{it} = \xi(X_{it}, \mathcal{A}_i, \mathcal{U}_{it}) + W_{it}'\gamma + \lambda_t$. Section A in the supplementary

[7] Note here that requiring stayer subpopulations implies that we cannot use variables, such as age, which cannot stay constant over time.

appendix contains the relevant identification and testing results.

An alternative approach to identifying average partial effects is to use restrictions on individual heterogeneity. The following theorem illustrates how the conditional random effects assumption fulfills the generalized first-differencing condition and also has a testable restriction for $T = 2$.

Theorem 2.2 ***Conditional Random Effects: Identification & Testable Restrictions** ($T = 2$)*

Let Assumption 2.1 hold.

$$\begin{aligned}
&\textit{If} \quad \mathcal{A}_i, \mathcal{U}_{i1}, \mathcal{U}_{i2} | X_i \overset{d}{=} \mathcal{A}_i, \mathcal{U}_{i1}, \mathcal{U}_{i2} | X_{i1}, \\
&\textit{then } (i) \int (\xi_2(x, a, u_2) - \xi_1(x, a, u_1)) \\
&\qquad \times \left(dF_{\mathcal{A}_i, \mathcal{U}_{i1}, \mathcal{U}_{i2}|X_i}(a, u_1, u_2 | (x, x')) - dF_{\mathcal{A}_i, \mathcal{U}_{i1}, \mathcal{U}_{i2}|X_i}(a, u_1, u_2 | (x, x)) \right) = 0 \\
&\qquad (ii)\ F_{Y_{i1}|X_i}(.|(x, x')) = F_{Y_{i1}|X_i}(.|(x, x)), \forall x, x' \in \mathbb{X}, x \neq x'
\end{aligned}$$

The above identifying assumption imposes an exclusion restriction on X_{i2} from the conditioning set of the distribution of all unobservables. One interpretation of this assumption is that selection is on the observables X_{i1}, and hence it relates to unconfoundedness. It is more appropriately referred to as a conditional random effects assumption. This identifying assumption is particularly suitable in field experiments, where pre- and post-treatment outcome variables are measured. The testable restriction in (ii) of the above theorem is inherently a testable restriction of random assignment of X_{i2}. If the treatment is in fact randomly assigned, then the distribution of the pre-treatment outcome variable in the first time period for the control group $X_i = (x, x)$ should be the same as that of the pre-treatment outcome variable of the treatment group $X_i = (x, x')$. Given the variety of implementation issues that may interfere with random assignment in field experiments, such as attrition and other selection problems, the above identifying assumption and the resulting testable restriction can be tailored to that setting as in the following example.

Example 1 ***Attrition in a Field Experiment***

Consider a simple example of attrition in a field experiment. Suppose that we have a field experiment with a single treatment, i.e. $X_{it}^{(1)} \in \{0,1\}$*, and we observe the control group* ($X_i^{(1)} = (0,0)$) *and treatment group* ($X_i^{(1)} = (0,1)$)*. The treatment is randomly assigned in the second period, hence we expect the treatment and control group to have the same distribution of unobservables. We observe the outcome variable before and after treatment,* Y_{i1} *and* Y_{i2}*, only for individuals which choose to respond pre- and post-treatment. Let* $X_i^{(2)}$ *be a binary variable for whether individuals respond or not. The identification question is whether* $E[Y_{i2}|X_i^{(1)} = (0,1), X_i^{(2)} = 1] - E[Y_{i2}|X_i^{(1)} = (0,0), X_i^{(2)} = 1]$ *can identify the APE* $\beta_2(0 \to 1|X_i^{(1)} = (0,1), X_i^{(2)} = 1)$*. Note that by the above theorem if the conditional random effects assumption holds, specifically* $\mathcal{A}_i, \mathcal{U}_{i1}, \mathcal{U}_{i2}|X_i^{(1)}, X_i^{(2)} \overset{d}{=} \mathcal{A}_i, \mathcal{U}_{i1}, \mathcal{U}_{i2}|X_{i1}^{(1)}, X_i^{(2)}$*,*[8] *then the APE in question is identified and we have the testable restriction,* $F_{Y_{i1}|X_i^{(1)},X_i^{(2)}}(.|(x,x'),1) = F_{Y_{i1}|X_i^{(1)},X_i^{(2)}}(.|(x,x),1)$*. The intuition here is that the identification of the APE is still possible if conditioning on whether individuals respond or not does not interfere with the random assignment of* $X_{i2}^{(1)}$*. The testable implication is hence that the pre-treatment outcome variable of the respondents in the control and treatment group has the same distribution.*

The exclusion restriction in Theorem 2.2 can alternatively be imposed on X_{i1} as opposed to X_{i2}, which yields $\mathcal{A}_i, \mathcal{U}_{i1}, \mathcal{U}_{i2}|X_i \overset{d}{=} \mathcal{A}_i, \mathcal{U}_{i1}, \mathcal{U}_{i2}|X_{i2}$.[9] This assumption allows X_{i2} to depend on past shocks $\mathcal{U}_{i1}$. In industrial organization (e.g. Olley and Pakes, 1996), $\mathcal{U}_{it}$ is a firm's productivity, which is unobservable to the econometrician, and $X_{i,t+1}$ are inputs. This variant of the conditional random effects assumption is more suitable here since it allows future inputs to depend on past productivity.

The conditional random effects assumption is a special case of correlated random effects assumptions, which have been considered in Altonji and Matzkin (2005) and Bester and Hansen (2009). Altonji and Matzkin (2005) shows how correlated random effects assumptions can achieve identification of APEs. The following theorem presents the testable restrictions for the general class of correlated random effects assumptions when $T \geq 2$. We first introduce

[8] Note that one can omit $X_i^{(1)}$ from the conditioning set in this example, since $X_{i1}^{(1)} = 0$ for all subpopulations in this example.

[9] Note that the testable restriction in this setup would be that $F_{Y_{i2}|X_i}(.|(x,x)) = F_{Y_{i2}|X_i}(.|(x',x))$ for all $x, x' \in \mathbb{X}$, $x \neq x'$.

the notation $\underline{x}_t$ to denote the t^{th} column of $\underline{x}$.

Theorem 2.3 ***Correlated Random Effects: Identification and Testable Restrictions ($T \geq 2$)***

Let Assumption 2.1 hold.

$$\begin{aligned}
&\textit{If} \quad \mathcal{A}_i, \mathcal{U}_{i1}, \mathcal{U}_{i2}, ..., \mathcal{U}_{iT} | X_i \overset{d}{=} \mathcal{A}_i, \mathcal{U}_{i1}, \mathcal{U}_{i2}, ..., \mathcal{U}_{iT} | h(X_i), \\
&\textit{then } (i) \beta_t(\underline{x}_t \to \underline{x}'_t | X_i = \underline{x}) = E[Y_{it} | X_i = \underline{x}'] - E[Y_{it} | X_i = \underline{x}], \textit{(Altonji and Matzkin, 2005)} \\
&\quad (ii) F_{Y_{i\tau}|X_i}(.|\underline{x}) = F_{Y_{i\tau}|X_i}(.|\underline{x}'), \\
&\quad \forall \underline{x}, \underline{x}' \in \mathbb{X}^T,\ t, \tau \in \{1, \dots, T\},\ \textit{where } \underline{x}_t \neq \underline{x}'_t,\ \underline{x}_\tau = \underline{x}'_\tau \textit{ and } h(\underline{x}) = h(\underline{x}'). \qquad (4)
\end{aligned}$$

In the above theorem, (i) is a re-statement of a result in Altonji and Matzkin (2005), whereas (ii) gives the testable restrictions for this class of identifying assumptions. Now we proceed to propose specification tests of all of the aforementioned restrictions.

3 Testing Identifying Assumptions

In this section, we develop statistics to test the restrictions implied by the identifying assumptions presented in Section 2. The APE of a subpopulation may be identified using different assumptions. For instance, Theorems 2.1 and 2.3 show how the APE can be identified using time homogeneity and correlated random effects, respectively. The statistics proposed here do not test whether the APE is identified or not. They test whether a particular identifying assumption holds. Their rejection is hence evidence against the assumption in question *only*. As pointed out above, the empirical setting typically suggests a suitable identification strategy.

The testable implications of the identifying assumptions we consider are equality restrictions on the conditional distribution of the outcome variable, hence they are an extension of the classical two-sample problem. The KS and CM statistics are well-known statistics for testing the equality of two distributions. Under the assumptions of the classical two-sample problem, i.e. the two samples are independent and each consists of i.i.d. observations of a continuous random variable, the two statistics have known asymptotic critical values that can be used in practice. There are three sources of differences between our setup and the classical two-sample problem. First, in cross-sectional panel data, we have dependence across time due to the time-invariant unobservables as well as time-series dependence of the idiosyncratic shocks. Secondly, to account for parallel trends in the case of time homogeneity, we have to compare the distribution of demeaned random variables. Thirdly, as illustrated in Theorem 2.1 and 2.2, each identifying assumption implies several equality restrictions that have to hold jointly. To obtain the p-value of all the statistics proposed below, we propose the following bootstrap procedure. Let T_n denote the statistic obtained from the original sample and T_n^b the properly centered statistic obtained from the b^{th} bootstrap sample. Let B denote the number of bootstrap replications.

Procedure 3.1 ***(Bootstrap Procedure)***

1. *Compute the statistic in question,* T_n*, for the original sample,* $\{\{Y_1, X_1\}, ..., \{Y_n, X_n\}\}$*.*
2. *Resample* n *observations* $\{\{\hat{Y}_1, \hat{X}_1\}, ..., \{\hat{Y}_n, \hat{X}_n\}\}$ *with replacement from the original sample. Compute* T_n^b*, the centered statistic for the* b^{th} *bootstrap sample .*
3. *Repeat 1-2* B *times.*
4. *Calculate the p-values of the tests with* $p_n = \sum_{b=1}^{B} 1\{T_n^b > T_n\}/B$*. Reject if p-value is smaller than some significance level* α*.*

The key feature that warrants some attention in the above procedure is that we are resampling individuals. We treat all of the observations of an individual as a single object. Hence, our procedure is valid without any restrictions on the time series dependence in our data. This is intuitive, since we have a fixed-T setup. The bootstrap procedure exploits cross-sectional independence. In the following, we show that our test statistics whose p-values are obtained using the above procedure are asymptotically valid to test time homogeneity and correlated random effects.

3.1 Testing Time Homogeneity

Theorem 2.1 establishes the testable restriction of time homogeneity for the $T = 2$ case. For $T > 2$, the testable restrictions are similarly given by the following

$$F_{Y_{i,t-1}|(X_{i,t-1},X_{it})}(.|(x,x)) = F_{Y_{it}-\Delta\lambda_t(x)|(X_{i,t-1},X_{it})}(.|(x,x)), \forall x \in \mathbb{X}, t = 2,\ldots,T, \quad (5)$$

where $\Delta\lambda_t(x) \equiv \lambda_t(x) - \lambda_{t-1}(x)$. Recall that $\mathbb{X}$ is the support of X_{it}, a $d_x \times 1$ vector, and $K \equiv |\mathbb{X}|$. The stayer subpopulations are the individuals that have the same value $x \in \mathbb{X}$ in both periods.[10] Hence, in the above we have $K \times (T-1)$ restrictions that we would like to test jointly.[11]

Let $\Delta X_{it} = X_{it} - X_{i,t-1}$. We can integrate (5) over $(X_{i,t-1}, X_{it}) = (x,x)$ conditional on $\Delta X_{it} = 0$ to obtain the following hypothesis

$$H_0^{gpt,T} : F_{Y_{i,t-1}|\Delta X_{it}}(.|0) = F_{Y_{it}|\Delta X_{it}}(.,\Delta\Lambda_t|0), \text{for } t = 2,\ldots,T, \quad (6)$$

where $\Delta\Lambda_t \equiv (\Delta\lambda_t(x^1), \Delta\lambda_t(x^2), \ldots, \Delta\lambda_t(x^K))$ and $F_{Y_{it}|\Delta X_{it}}(.,\Delta\Lambda_t|0) \equiv \sum_{k=1}^K P((X_{i,t-1}, X_{it}) = (x^k, x^k)|\Delta X_{it} = 0) F_{Y_{it}-\Delta\lambda_t(x^k)|(X_{i,t-1},X_{it})}(.|(x^k,x^k))$. The null hypothesis $H_0^{gpt,T}$ is an implication of the more disaggregated restriction in (5). While it controls size, there are power trade-offs that might have implications for finite-sample properties. This issue is explored through a simulation study in Section C in the supplementary appendix.

$H_0^{gpt,T}$ consists of $T-1$ restrictions. Thus, the following statistics test all $T-1$ restrictions

[10] For instance, if X_{it} includes two binary regressors for union and high school completion, then $K = 4$, since the elements in the support of X_{it} represent the four categories: union members with and without high school degrees as well as non-union members with and without high school degrees. Stayer subpopulations are individuals that remained in one of these four categories in two time periods. In general, K is equal to the number of dummy variables in a fully saturated model with discrete regressors as described in Angrist (2001).

[11] The restriction in (5) is a natural extension of the testable restriction in Theorem 2.1 for the $T = 2$ case. It is possible however to write a more disaggregated hypothesis. Note that $F_{Y_{it}|(X_{i,t-1},X_{i,t})}(y|(x,x)) = \sum_{\underline{x}\in\mathbb{X}_{x,t}^T} P(X_i = \underline{x}|(X_{i,t-1}, X_{it}) = (x,x)) F_{Y_{it}|X_i}(y|\underline{x})$, where $\mathbb{X}_{x,t}^T = \{\underline{x} \in \mathbb{X}^T : (\underline{x}_{t-1}, \underline{x}_t) = (x,x)\}$ for $x \in \mathbb{X}$ and $t = 2,\ldots,T$. Thus, the testable restriction can be alternatively written as $F_{Y_{i,t-1}|X_i}(.|\underline{x}) = F_{Y_{it}-\Delta\lambda_t(x)|X_i}(.|\underline{x})$, $\forall \underline{x} \in \mathbb{X}_{x,t}^T, x \in \mathbb{X}, t = 2,...,T$.

jointly,

$$KS_{n,\mathbb{Y}}^{gpt,T} = \frac{1}{T-1}\sum_{t=2}^{T}\left\|\sqrt{n}\left\{F_{n,Y_{i,t-1}|\Delta X_{it}}(.|0) - F_{n,Y_{it}|\Delta X_{it}}(.,\Delta\Lambda_{n,t}|0)\right\}\right\|_{\infty,\mathbb{Y}},$$
$$CM_{n,\phi}^{gpt,T} = \frac{1}{T-1}\sum_{t=2}^{T}\left\|\sqrt{n}\left\{F_{n,Y_{i,t-1}|\Delta X_{it}}(.|0) - F_{n,Y_{it}|\Delta X_{it}}(.,\Delta\Lambda_{n,t}|0)\right\}\right\|_{2,\phi}, \quad (7)$$

where $F_{n,Y_{it}|\Delta X_{it}}$ and $\Delta\Lambda_{n,t}$ denote sample analogues of $F_{Y_{it}|\Delta X_{it}}$ and $\Delta\Lambda_t$, respectively. Procedure 3.1 can be used to obtain p-values for the above statistics with the following centered statistics for the bootstrap sample

$$\begin{aligned}KS_{n,\mathbb{Y}}^{gpt,T,b} = \frac{1}{T-1}\sum_{t=2}^{T}\Big\|&\sqrt{n}\left\{F^b_{n,Y_{i,t-1}|\Delta X_{it}}(.|0) - F^b_{n,Y_{it}|\Delta X_{it}}(.,\Delta\Lambda^b_{n,t}|0)\right\}\\ &- \sqrt{n}\left\{F_{n,Y_{i,t-1}|\Delta X_{it}}(.|0) - F_{n,Y_{it}|\Delta X_{it}}(.,\Delta\Lambda_{n,t}|0)\right\}\Big\|_{\infty,\mathbb{Y}},\\ CM_{n,\phi}^{gpt,T,b} = \frac{1}{T-1}\sum_{t=2}^{T}\Big\|&\sqrt{n}\left\{F^b_{n,Y_{i,t-1}|\Delta X_{it}}(.|0) - F^b_{n,Y_{it}|\Delta X_{it}}(.,\Delta\Lambda^b_{n,t}|0)\right\}\\ &- \sqrt{n}\left\{F_{n,Y_{i,t-1}|\Delta X_{it}}(.|0) - F_{n,Y_{it}|\Delta X_{it}}(.,\Delta\Lambda_{n,t}|0)\right\}\Big\|_{2,\phi},\end{aligned}$$

where $F^b_{n,Y_{it}|\Delta X_{it}}$ and $\Delta\Lambda^b_{n,t}$ denote bootstrap sample analogues of $F_{Y_{it}|\Delta X_{it}}$ and $\Delta\Lambda_t$, respectively, for the b^{th} bootstrap sample.

The following theorem shows that the bootstrap critical values of the above test statistics deliver correct asymptotic size and consistency against fixed alternatives. First, we have to impose an additional assumption.

Assumption 3.1 *(Bounded Density)*

$F_{Y_{it}}(.)$ has a density $f_{Y_{it}}(.)$ that is bounded, i.e. $\sup_{y\in\mathbb{Y}}|f_{Y_{it}}(y)| < \infty$, $t = 1, \ldots, T$.

Theorem 3.1 *Given that $\{Y_i, X_i\}_{i=1}^n$ is an iid sequence, $|\mathbb{X}| = K$, $P(\Delta X_{it} = 0) > 0$ for $t = 2, \ldots, T$, $F_{Y_{it}|X_i}(.|\underline{x})$ is non-degenerate for $t = 1, \ldots, T$ and $\underline{x} \in \mathbb{X}^T$, and Assumption 3.1 holds. Procedure 3.1 for $KS_{n,\mathbb{Y}}^{gpt,T}$ and $CM_{n,\phi}^{gpt,T}$ to test $H_0^{gpt,T}$ (i) provides correct asymptotic size α and (ii) is consistent against any fixed alternative of $H_0^{gpt,T}$.*

The proof is given in Appendix A.2. Assumption 3.1 merits some discussion. By demeaning the variables, we are introducing asymptotically normal noise to the empirical process.

Assumption 3.1 ensures that the empirical process converges nonetheless to a Brownian bridge by allowing us to apply the functional delta method. From here, it is straightforward to show that the bootstrap empirical process converges to the same tight limit process as the empirical process. Then, we show that the bootstrap critical values of the test statistics deliver correct asymptotic size and are consistent against fixed alternatives.

Alternatively, one could approach the $T > 2$ case as a multiple testing problem, where we have $T-1$ hypotheses of time homogeneity. For $t = 2, \dots, T$,

$$H_0^{gpt,t} : F_{Y_{i,t-1}|\Delta X_{it}}(.|0) = F_{Y_{it}|\Delta X_{it}}(., \Delta\Lambda_t|0).$$

For each t, the statistics can be computed similar to the $T = 2$ case. A multiple-testing correction procedure, such as the step-down procedures in Romano and Wolf (2005) and Romano, Shaikh, and Wolf (2008), can then be adapted to control the family-wise error rate for testing $\{H_0^{gpt,t}\}_{t=2}^T$.

3.2 Testing Correlated Random Effects Assumptions

Theorem 2.3 gives the testable restriction of the correlated random effects assumption as follows

$$F_{Y_{it}|X_i}(.|\underline{x}) = F_{Y_{it}|X_i}(.|\underline{x}'), \forall \underline{x}, \underline{x}' \in \mathbb{X}, \forall t = 1, 2, \dots, T, \text{ where } \underline{x}_t = \underline{x}'_t,\ h(\underline{x}) = h(\underline{x}'). \tag{8}$$

$h : \mathbb{X}^T \mapsto \mathbb{H}$ is the restriction imposed on the distribution of unobservables. For $T = 2$ and $h(X_i) = X_{i1}$, this delivers the conditional random effects assumption as in Theorem 2.2. Note that the restriction implies that $L \equiv |\mathbb{H}| < |\mathbb{X}^T|$, where $\mathbb{H} = \{h_l\}_{l=1}^L$.[12]

Now we introduce some notation in order to write the testable restriction. For $l = 1, 2, \dots, L$, define $\mathbb{X}_{t,k,l}^T \equiv \{\underline{x} \in \mathbb{X}^T : \underline{x}_t = x^k, h(\underline{x}) = h_l\}$. This is the set of subpopulations that have the same realization of X_{it} at time t, x^k, and have the same distribution of

[12] Correlated random effects strategies restrict heterogeneity across subpopulations. Two subpopulations, $\underline{x}$ and $\underline{x}'$, will have the same distribution of unobservables if $h(\underline{x}) = h(\underline{x}')$. Thus, it follows that $|\mathbb{H}| < |\mathbb{X}^T|$.

unobservables, since $h(\underline{x}) = h_l$. The testable restrictions state that the distribution of the outcome variable at time t for all subpopulations in this set are expected to be equal. This holds for all time periods and all K elements in $\mathbb{X}$, which yields the following hypothesis,

$$H_0^{cre} : F_{Y_{it}|X_i}(.|\underline{x}) = F_{Y_{it}|X_i}(.|\underline{x}'),$$
$$\forall \underline{x}, \underline{x}' \in \mathbb{X}^T_{k,l,t},\ k = 1, 2, \ldots, K,\ l = 1, 2, \ldots, L,\ t = 1, 2, \ldots, T.$$

For the conditional random effects assumption and $T = 2$, $\underline{x} = (x^k, x^k)$ and $\underline{x}' = (x^k, x^j)$ $j \neq k$, $j, k = 1, 2, \ldots, K$. The testable restrictions apply for $t = 1$ only. In that case, H_0^{cre} simplifies to the testable restriction in Theorem 2.2.

When $|\mathbb{X}^T_{k,l,t}| > 2$, we have equality restrictions on more than two cdfs. Hence, it is a multiple sample problem. In this setting, Quessy and Ethier (2012) introduce an "averaged" cdf. We follow their approach here and introduce the following cdf,

$$\bar{F}_{k,l,t} = \frac{1}{|\mathbb{X}^T_{k,l,t}|} \sum_{\underline{x} \in \mathbb{X}^T_{k,l,t}} F_{Y_{it}|X_i}(.|\underline{x}), \tag{9}$$

which averages over the cdfs of the outcome variable at time period t for subpopulations in the set $\mathbb{X}^T_{k,l,t}$.

The KS and CM statistics for the restrictions in time period t and fixing x^k are given by

$$KS^{cre}_{n,k,t} = \sum_{l=1}^{L} P_n(h(X_i) = h_l) \sum_{\underline{x} \in \mathbb{X}^T_{k,l,t}} P_n(X_i = \underline{x}|h(X_i) = h_l) \|\sqrt{n}\{F_{n,Y_{it}|X_i}(.|\underline{x}) - \bar{F}_{n,k,l,t}(.)\}\|_{\infty,\mathbb{Y}}$$
$$CM^{cre}_{n,k,t} = \sum_{l=1}^{L} P_n(h(X_i) = h_l) \sum_{\underline{x} \in \mathbb{X}^T_{k,l,t}} P_n(X_i = \underline{x}|h(X_i) = h_l) \|\sqrt{n}\{F_{n,Y_{it}|X_i}(.|\underline{x}) - \bar{F}_{n,k,l,t}(.)\}\|_{2,\phi}$$

Averaging the above statistics over k and t, we obtain the statistics that test H_0^{cre},

$$KS^{cre}_{n,\mathbb{Y}} = \frac{1}{KT} \sum_{k=1}^{K} \sum_{t=1}^{T} KS^{cre}_{n,k,t},$$
$$CM^{cre}_{n,\phi} = \frac{1}{KT} \sum_{k=1}^{K} \sum_{t=1}^{T} CM^{cre}_{n,k,t}.$$

The bootstrap-centered statistics are given by $KS^{cre,b}_{n,\mathbb{Y}} = \sum_{k=1}^{K}\sum_{t=1}^{T} KS^{cre,b}_{n,k,t}/KT$ and $CM^{cre,b}_{n,\phi} = \sum_{k=1}^{K}\sum_{t=1}^{T} CM^{cre,b}_{n,k,t}/KT$, where

$$\begin{aligned}
KS^{cre,b}_{n,k,t} &= \sum_{l=1}^{L} P^b_n(h(X_i)=h_l) \sum_{\underline{x}\in\mathbb{X}^T_{k,l,t}} P^b_n(X_i=\underline{x}|h(X_i)=h_l) \\
&\quad \times \|\sqrt{n}\{F^b_{n,Y_{it}|X_i}(.|\underline{x}) - \bar{F}^b_{n,k,l,t}(.) - (F_{n,Y_{it}|X_i}(.|\underline{x}) - \bar{F}_{n,k,l,t}(.))\}\|_{\infty,\mathbb{Y}}, \\
CM^{cre,b}_{n,k,t} &= \sum_{l=1}^{L} P^b_n(h(X_i)=h_l) \sum_{\underline{x}\in\mathbb{X}^T_{k,l,t}} P^b_n(X_i=\underline{x}|h(X_i)=h_l) \\
&\quad \times \|\sqrt{n}\{F^b_{n,Y_{it}|X_i}(.|\underline{x}) - \bar{F}^b_{n,k,l,t}(.) - (F_{n,Y_{it}|X_i}(.|\underline{x}) - \bar{F}_{n,k,l,t}(.))\}\|_{2,\phi}.
\end{aligned}$$

The above statistics use the bootstrap empirical probabilities $P^b_n(A)$ for an event A.

The following theorem shows that the bootstrap critical values for the above statistics obtained using Procedure 3.1 are asymptotically valid.

Theorem 3.2 *Given $\{Y_i, X_i\}_{i=1}^n$ is an iid sequence, $|\mathbb{X}| = K$, $P(X_i = \underline{x}) > 0$ for all $\underline{x} \in \mathbb{X}^T$, and $F_{Y_{it}|X_i}(.)$ is non-degenerate for all $t = 1, 2, \ldots, T$, the procedure described in 1-4 for $KS^{cre}_{n,\mathbb{Y}}$ and $CM^{cre}_{n,\phi}$ to test H^{cre}_0 (i) provides correct asymptotic size α and (ii) is consistent against fixed alternatives.*

The proof is in Appendix A.2. The convergence of the empirical and bootstrap empirical processes to a tight Brownian bridge follows from results in Van der Vaart and Wellner (2000). The remainder of the proof follows by similar arguments to Theorem 3.1.

3.3 Monte Carlo Study

In this section, we examine the finite-sample performance of the proposed bootstrap procedure for obtaining the critical values of the KS and CM statistics to test time homogeneity and conditional random effects assumptions. For time homogeneity, we will consider several of its variants, time homogeneity with no trend ($\lambda_t(x) = 0$), a parallel trend ($\lambda_t(x) = \lambda_t$ for all $x \in \mathbb{X}$), and a generalized parallel trend ($\lambda_t(x)$).

Table 1 describes the models we consider in our Monte Carlo design. The structural function $\xi(x, a, u)$ is adapted from the design in Evdokimov (2010) to include a location shift

μ_0 in order to maintain $E[Y_{i1}]$ to be the same for all models A-D.[13] X_{it} is a binomial random variable that is standardized to have mean 0 and standard deviation 1. Models A-C are variants of time homogeneity. Model A exhibits time homogeneity without a parallel trend. Model B allows for a parallel trend (λ_t). Model C allows for a generalized parallel trend that depends on the regressor ($\lambda_t \text{sign}(X_{it})$). If $\lambda_t = \lambda_1$ for all $t = 2, \ldots, T$, Models A, B and C are all equivalent. Thus, $\lambda_t - \lambda_1$ exhibits the location shift by which Models B and C deviate from Model A. The difference between Models B and C is that the latter allows subpopulations with positive X_{it} to have a location shift equal to λ_t, whereas subpopulations with negative X_{it} to have a location shift equal to $-\lambda_t$. Model D exhibits the conditional random effects assumption, while allowing for a time-varying scale shift in the structural function, σ_t. If $\sigma_t = \sigma_1$ for all $t = 2, \ldots, T$, then Model D also exhibits time homogeneity with a parallel trend, λ_t. Thus, Model D deviates from time homogeneity with a parallel trend (Model B) by $\sigma_t - \sigma_1$. In the simulations, we impose the normalizations, $\lambda_1 = 0$ and $\sigma_1 = 1$.

This section is organized as follows. Section 3.3.1 presents the baseline Monte Carlo results for testing time homogeneity and conditional random effects. Section 3.3.2 examines the behavior of the statistics of time homogeneity with a parallel trend and a generalized parallel trend in a design resembling the National Longitudinal Survey of Youth (NLSY) subsample which we consider in the empirical illustration. Sections C.1-C.3 of the supplementary appendix includes additional simulation results that examine the choice of density in the CM statistic and the relative performance of aggregated and disaggregated statistics for testing time homogeneity. The simulation results point out several issues that practitioners have to address in implementing the tests proposed here. Section C.4 of the supplementary appendix summarizes the resulting recommendations for practitioners.

3.3.1 Testing Identifying Assumptions: Baseline Results

In the baseline results, we examine the finite-sample behavior of the bootstrap procedure for $n = 500, 2000$, $T = K = 2$, and $p = 0.5$. Thus, we expect half of the sample to be stayers

[13] This point relates to the choice of density used to compute the CM statistic and is discussed in greater detail in Section C.1 in the supplementary appendix.

Table 1: Models Considered in the Monte Carlo Design

Model	$\mathcal{A}_i$	$\mathcal{U}_{it}$	Y_{it}
A	$\mathcal{A}_i = 0.5\sqrt{T}\bar{X}_T + 0.5\psi_i$	$\mathcal{U}_{it} = \epsilon_{it}\bar{X}_T$	$Y_{it} = \xi(X_{it}, \mathcal{A}_i, \mathcal{U}_{it})$
B	"	"	$Y_{it} = \xi(X_{it}, \mathcal{A}_i, \mathcal{U}_{it}) + \lambda_t$
C	"	"	$Y_{it} = \xi(X_{it}, \mathcal{A}_i, \mathcal{U}_{it}) + \lambda_t \text{sign}(X_{it})$
D	$\mathcal{A}_i = 0.5\sqrt{T}X_{i1} + 0.5\psi_i$	$\mathcal{U}_{it} = X_{i1}\epsilon_{it}$	$Y_{it} = \xi(X_{it}, \mathcal{A}_i, \mathcal{U}_{it})\sigma_t + \lambda_t$

Notes: $\xi(x, a, u) = \mu_0 + a + (2 + a)x + u$; $X_{it} = \{Z_{it} - p(K-1)\}/\sqrt{(K-1)p(1-p)}$, $Z_{it} \overset{i.i.d.}{\sim}$ Binomial(K-1, p); $\epsilon_{it} \overset{i.i.d.}{\sim}$ N(0, 1); $\psi_i \overset{i.i.d.}{\sim}$ N(0,1); $\bar{X}_T \equiv \sum_{t=1}^{T} X_{it}/T$; sign$(g(x))$ $= 1\{g(x) \geq 0\} - 1\{g(x) < 0\}$; $\lambda_1 = 0$, $\sigma_1 = 1$.

and half to be movers. In this design, we set $E[Y_{i1}] = 0$ by assigning μ_0 an appropriate value.[14] In terms of the models that exhibit time homogeneity, we consider Model A as well as Models B and C, with $\lambda_2 = 0.25$, which is about 10% of the standard deviation of Y_{i1}.[15] For Model D, we set $\lambda_2 = 0.5$ and $\sigma_2 = 1.1$.

Under each model, we compute the bootstrap-adjusted p-values using Procedure 3.1 with $B = 200$ for the following statistics:

$$KS^{nt}_{n,\mathbb{Y}} = \|F_{n,Y_{i1}|\Delta X_{i2}}(.|0) - F_{n,Y_{i2}|\Delta X_{i2}}(.|0)\|_{\infty,\mathbb{Y}} \quad (10)$$

$$KS^{pt}_{n,\mathbb{Y}} = \|F_{n,Y_{i1}|\Delta X_{i2}}(.|0) - F_{n,Y_{i2}-\Delta\lambda_n|\Delta X_{i2}}(.|0)\|_{\infty,\mathbb{Y}} \quad (11)$$

$$KS^{gpt}_{n,\mathbb{Y}} = \|F_{n,Y_{i1}|\Delta X_{i2}}(.|0) - F_{n,Y_{i2}|\Delta X_{i2}}(., \Delta\Lambda_n|0)\|_{\infty,\mathbb{Y}} \quad (12)$$

$$KS^{cre}_{n,\mathbb{Y}} = \sum_{l=1}^{K} P_n(X_{i1} = x^l)\frac{1}{K}\sum_{k=1}^{K}\|F_{n,Y_{i1}|X_i}(.|(x^l, x^k)) - \bar{F}_{n,l,1}(.)\|_{\infty,\mathbb{Y}} \quad (13)$$

$$CM^{nt}_{n,\phi} = \|F_{n,Y_{i1}|\Delta X_{i2}}(.|0) - F_{n,Y_{i2}|\Delta X_{i2}}(.|0)\|_{2,\phi} \quad (14)$$

$$CM^{pt}_{n,\phi} = \|F_{n,Y_{i1}|\Delta X_{i2}}(.|0) - F_{n,Y_{i2}-\Delta\lambda_n|\Delta X_{i2}}(.|0)\|_{2,\phi} \quad (15)$$

$$CM^{gpt}_{n,\phi} = \|F_{n,Y_{i1}|\Delta X_{i2}}(.|0) - F_{n,Y_{i2}|\Delta X_{i2}}(., \Delta\Lambda_n|0)\|_{2,\phi} \quad (16)$$

$$CM^{cre}_{n,\phi} = \sum_{l=1}^{K} P_n(X_{i1} = x^l)\frac{1}{K}\sum_{k=1}^{K}\|F_{n,Y_{i1}|X_i}(.|(x^l, x^k)) - \bar{F}_{n,l,1}(.)\|_{2,\phi}. \quad (17)$$

where $\bar{F}_{n,l,1}(.) = \sum_{k=1}^{K} F_{n,Y_{i1}|X_i}(.|(l, x^k))/K$, $\Delta\lambda_n = \sum_{i=1}^{n}(Y_{i2} - Y_{i1})1\{\Delta X_{i2} = 0\}/\sum_{i=1}^{n} 1\{\Delta X_{i2} =$

[14] We numerically calculate μ_0 by calculating the $E[Y_{i1}]$ for models A-D, where $\xi(x, a, u) = a + (2 + a)x + u$, using a sample of size $10,000,000$. For models A-C, $\mu_0 = 0.354$. For model D, $\mu_0 = 0.707$.

[15] The standard deviation of Y_{i1} is numerically calculated to be 2.6, using a sample of size $10,000,000$.

$0\}$. $\Delta\Lambda_n$ is defined in Section 3.1. In our baseline study, we set ϕ to be the standard normal density. We examine the behavior of the CM statistic using other densities in Section C.1 of the supplementary appendix. Both the KS and CM statistics are computed on a grid $\{\underline{y}, \underline{y}+0.01, \underline{y}+0.02, \ldots, \bar{y}\}$, where $\underline{y} = \min_{i,t} \tilde{Y}_{it}$, $\bar{y} = \max_{i,t} \tilde{Y}_{it}$, and $\tilde{Y}_{it}$ denotes the appropriately demeaned Y_{it}.[16] Note that the statistics with nt super-script test time homogeneity with no trend, pt test time homogeneity with a parallel trend, gpt test time homogeneity with a generalized parallel trend, $\lambda_t(X_{it})$, and cre test the conditional random effects assumption. For $K = 2$, $\mathbb{X} = \{-1, 1\}$ and the cre KS statistic simplifies to the following,

$$\begin{aligned} KS^{cre}_{n,\mathbb{Y}} = &P_n(X_{i1} = -1)\|F_{n,Y_{i1}|X_i}(.|(-1,-1)) - F_{n,Y_{i1}|X_i}(.|(-1,1))\|_{\infty,\mathbb{Y}} \\ &+ P_n(X_{i1} = 1)\|F_{n,Y_{i1}|X_i}(.|(1,-1)) - F_{n,Y_{i1}|X_i}(.|(1,1))\|_{\infty,\mathbb{Y}}. \end{aligned}$$

The respective CM statistic follows by substituting $\|.\|_{2,\phi}$ in lieu of $\|.\|_{\infty,\mathbb{Y}}$ in the above.

Table 2 reports the rejection probabilities using the bootstrap-adjusted p-values for the above statistics in the baseline Monte Carlo design. Under Models A-C, we find that the finite-sample behavior of the bootstrap-adjusted KS and CM statistics are very similar. They both provide good size control and have fairly similar power properties. As expected, finite-sample power improves as n increases. It is worth noting that for these models, the KS statistic for pt and gpt, especially for $n = 500$, tends to be under-sized.

Under Model D, the cre statistics provide good size control, where the CM statistic seems to fare better in this regard than the KS statistic. However, for pt, the KS statistic exhibits better finite-sample power than the CM statistic, whereas the latter has better power properties than the former for gpt. Overall, the gpt statistics seem less powerful in detecting time heterogeneity in the scale as in Model D compared to pt. This may be due to two factors. First, the gpt allows for parallel trends that depend on the regressor. Some of the time heterogeneity in the scale may be mistakenly soaked up by these generalized parallel trends, which makes the gpt statistics less powerful at detecting scale deviations from time homogeneity relative to the pt statistics. Secondly, the pt uses an estimate of λ_2 that uses

[16]For instance for the pt statistics, $\tilde{Y}_{i1} = Y_{i1}$ and $\tilde{Y}_{i2} = Y_{i2} - \Delta\lambda_n$.

the entire stayer subsample, which is expected to be $n/2$ in our design, whereas the *gpt* uses estimates of $\lambda_2(-1)$ and $\lambda_2(1)$ for subsamples $X_i = \underline{x} \in \{(-1,-1),(1,1)\}$, respectively. The size of these subsamples is expected to be $n/4$. Hence, these estimates will be noisier than the estimates of λ_2 used to construct the *pt* statistics, which may affect the finite-sample power properties of the *gpt* statistics.

3.3.2 Monte Carlo Study Resembling NLSY 1983-1987

In this section, we design a Monte Carlo experiment that resembles the subsample of the NLSY 1983-1987, which we will use in our empirical illustration. Thus, we set $n = 1000$, $T = 5$, and $K = 16$. The schooling variable, measured by the highest grade completed, takes integer values between 6 and 20 in the NLSY sample, hence our choice of K for the simulations. We adjust the models in Table 1 in order to match the mean and standard deviation of the outcome variable, log hourly wage, in the NLSY sample we consider. There are two main adjustments:

(1) The regressors are generated to increase the proportion of stayers to resemble the proportion of stayers in the NLSY. X_{it} is defined as in the design in Table 1. However, we change the design of Z_{it} as follows

$$
\begin{aligned}
Z_{i1} &\overset{i.i.d.}{\sim} Bin(K-1, p) \\
Z_{i2} &= Z_{i1} + 1\{\pi_{i2} > 1\}1\{Z_{i1} < K-1\} \\
Z_{i3} &= Z_{i2} + 1\{\pi_{i3} > 1.5\}1\{Z_{i2} < K-1\} \\
Z_{i4} &= Z_{i3} + 1\{\pi_{i4} > 2\}1\{Z_{i3} < K-1\} \\
Z_{i5} &= Z_{i4} + 1\{\pi_{i5} > 2\}1\{Z_{i4} < K-1\}
\end{aligned}
$$

where $\pi_{it} \overset{i.i.d.}{\sim} N(0,1)$ across i, t.

(2) The structural function used in this design is given by: $\xi(x,a,u) = \mu_0 + (a + (2+a)x + u)/c_0$, where we choose μ_0 and c_0 so that the means and standard deviations of the outcome variable in each time period of the design matches the annual means and standard

deviations of log hourly wage in the NLSY 1983-1987, which are reported in Table 5.[17]

It remains to set the values for λ_t and σ_t for $t = 1, \ldots, 5$. For Models B and C, we set $\lambda_1 = 0$ and fix $\lambda_{t+1} - \lambda_t = a_t$, where we let $a_t = 0.005, 0.01, 0.025$, which correspond to 0.01, 0.02 and 0.05 proportion of the standard deviation of Y_{i1}, respectively.[18] For Model D, we fix $a_t = 0.01$. As for σ_t, we set $(\sigma_1, \sigma_2, \sigma_3, \sigma_4, \sigma_5) = (1, 1 + b_2, 1, 1 - b_2, 1)$, where we let $b_2 = 0.025, 0.05, 0.1$, which correspond to proportions of the standard deviation of Y_{i1}. Table 3 reports the proportion of the stayer subpopulation relative to the entire population for every two periods, $P(X_{i,t-1} = X_{it})$, as well as population means and standard deviations of Y_{it} for all time periods for some of the models we consider here. All quantities are numerically calculated using a sample of size $n = 10,000,000$. Comparing these quantities with their corresponding quantities for the NLSY subsample in 1983-1987 in Table 5, we find that the models we consider in our simulation study seem to match the data quite well in terms of the year-to-year proportion of the stayers to the entire sample as well as the mean and standard deviation of log hourly wage.

In this section, we focus on the tests of time homogeneity with a parallel trend (pt) and a generalized parallel trend (gpt). Since $T = 5$, we use the following statistics,

$$KS_{n,\mathbb{Y}}^{pt,T} = \frac{1}{T-1} \sum_{t=2}^{T} \left\| \sqrt{n} \left\{ F_{n,Y_{i,t-1}|\Delta X_{it}}(.|0) - F_{n,Y_{it}-\Delta\lambda_{n,t}|\Delta X_{it}}(.|0) \right\} \right\|_{\infty,\mathbb{Y}}, \quad (18)$$

$$KS_{n,\mathbb{Y}}^{gpt,T} = \frac{1}{T-1} \sum_{t=2}^{T} \left\| \sqrt{n} \left\{ F_{n,Y_{i,t-1}|\Delta X_{it}}(.|0) - F_{n,Y_{it}|\Delta X_{it}}(., \Delta\Lambda_{n,t}|0) \right\} \right\|_{\infty,\mathbb{Y}}, \quad (19)$$

$$CM_{n,\phi}^{pt,T} = \frac{1}{T-1} \sum_{t=2}^{T} \left\| \sqrt{n} \left\{ F_{n,Y_{i,t-1}|\Delta X_{it}}(.|0) - F_{n,Y_{it}-\Delta\lambda_{n,t}|\Delta X_{it}}(.|0) \right\} \right\|_{2,\phi}, \quad (20)$$

$$CM_{n,\phi}^{gpt,T} = \frac{1}{T-1} \sum_{t=2}^{T} \left\| \sqrt{n} \left\{ F_{n,Y_{i,t-1}|\Delta X_{it}}(.|0) - F_{n,Y_{it}|\Delta X_{it}}(., \Delta\Lambda_{n,t}|0) \right\} \right\|_{2,\phi}, \quad (21)$$

[17]For Models A-C, $\mu_0 = 5.95$; for Model D, $\mu_0 = 6.14$. For Models A-C, $c_0 = 2 \times 2.61$; for Model D, $c_0 = 2 \times 3.77$. The quantities for μ_0 and $c_0/2$ are numerically calculated from the expected value and standard deviation of Y_{i1}, respectively, using a sample of size $10,000,000$, where $\xi(x, a, u) = a + (2 + a)x + u$.

[18]It is worth noting that under Model B, $a_t = 0.025$ corresponds to a 2.5% average increase from year-to-year in the outcome variable for stayer subpopulations. Under Model C, it would correspond to a 5% difference in average increase in the outcome variable between stayer subpopulations with above average X_{it} and those below average X_{it}.

where $\Delta\lambda_{n,t} = \sum_{i=1}^n \Delta Y_{it} 1\{\Delta X_{it} = 0\} / \sum_{i=1}^n 1\{\Delta X_{it} = 0\}$, $\Delta\lambda_{n,t}(x) = \sum_{i=1}^n \Delta Y_{it} 1\{(X_{i,t-1}, X_{it}) = (x,x)\} / \sum_{i=1}^n 1\{(X_{i,t-1}, X_{it}) = (x,x)\}$, $\Delta\Lambda_{n,t} = \left(\Delta\lambda_{n,t}(x^1), \ldots, \Delta\lambda_{n,t}(x^K)\right)'$. We consider the CM statistics with different choices of ϕ: $N(6.5, 0.25)$, $N(6.5, 0.5)$ and $U(0, 14)$. The mean of the normal densities is chosen to be close to the overall mean of Y_{it} across individuals and time. The standard deviation of Y_{it} is about 0.5, thus we consider two normal densities, one with the same and one with a smaller standard deviation than the outcome variable. The uniform density we consider gives equal weight to a large proportion of the support of Y_{it}.

Table 4 reports the simulation rejection probabilities for the above statistics using the bootstrap-adjusted p-values. Overall, the KS and CM statistics for *pt* control size for Models A and B. They also reflect good finite-sample power properties under Models C and D. The CM statistic with $N(6.5, 0.25)$ performs better than with $N(6.5, 0.5)$ under small deviations from the *pt* null hypothesis, i.e. Model C with $\lambda_{t+1} - \lambda_t = 0.01$ and Model D $|\sigma_t - \sigma_{t+1}| = 0.025, 0.05$, which suggests that giving higher weight to the center of the distribution improves finite-sample power in our setup. The CM statistic with $U(0, 14)$ seems to be fairly close to the performance of the CM statistic with $N(6.5, 0.25)$. The *gpt* statistics tend to be quite under-sized under Models A-C and has no power under Model D in our simulation study. Section C.3 in the supplementary appendix proposes alternative CM statistics that perform substantially better for the *gpt* hypothesis in simulations.

4 Empirical Illustration: Returns to Schooling

4.1 Revisiting Angrist and Newey (1991)

Consider the linear fixed effects model where for $i = 1, \ldots, n$, $t = 1, \ldots, T$

$$Y_{it} = X_{it}'\beta + \mathcal{A}_i + \mathcal{U}_{it}, \tag{22}$$

$$E[\mathcal{U}_{it} | X_{i1}, \ldots, X_{iT}, \mathcal{A}_i] = 0 \tag{23}$$

where X_{it} and β are $d_x \times 1$ vectors. The key identifying assumption in the above is the restriction in (23), which is referred to by Chamberlain (1984) as "strict exogeneity conditional on a latent variable." Chamberlain (1984) proposes a minimum chi-square (MCS) procedure to test the restrictions implied by (22) and (23). Angrist and Newey (1991) show that the over-identification tests from three stage least squares (3SLS) equivalents of the Chamberlain (1984) procedure yield statistics that are identical to the MCS statistic.[19] The testable restrictions are obtained from considering the linear prediction of $\mathcal{A}_i$ given X_i

$$\mathcal{A}_i = X_{i1}'\eta_1 + \cdots + X_{iT}'\eta_T + \mathcal{E}_i. \tag{24}$$

By construction, $\mathcal{E}_i$ is uncorrelated with $X_{i1}, \ldots, X_{iT}$.[20] Taking deviations of Y_{it} from Y_{iT} for $t = 1, \ldots, T-1$ and plugging (24) into Y_{iT} yields the equations for the simplest 3SLS procedure

$$\begin{aligned} Y_{i1} - Y_{iT} &= (X_{i1} - X_{iT})'\beta + (\mathcal{U}_{i1} - \mathcal{U}_{iT}) \\ &\vdots \\ Y_{i1} - Y_{i,T-1} &= (X_{i,T-1} - X_{iT})'\beta + (\mathcal{U}_{i,T-1} - \mathcal{U}_{iT}) \\ Y_{iT} &= X_{iT}'\beta + X_{i1}'\eta_1 + \cdots + X_{iT}'\eta_T + \mathcal{E}_i + \mathcal{U}_{iT}. \end{aligned}$$

Using a subsample of the national longitudinal survey of youth (NLSY) 1983-1987, Angrist and Newey (1991) estimate the union-wage effects as well as returns to schooling and apply their over-identification test.[21] Their over-identification test does not reject for the union-wage equation, but rejects for the returns to schooling equation at the 5% level. In Section 4.2, we will revisit the returns to schooling application and apply the test for time homogeneity in the presence of a parallel trend, which is a test of nonparametric identification.

[19] They also show that it can simplify to the degrees of freedom times R^2 from the regression of the residuals from the analysis of covariance (ANACOVA) on all leads and lags of the regressors when some transformations of the time-varying unobservables are homoskedastic.

[20] Chamberlain (1984) points out that (24) is not restrictive under the assumptions that variances are finite and that the joint distribution of $(X_i, \mathcal{A}_i)$ does not depend on i.

[21] Returns to schooling is not usually identified in the panel data context, because individual schooling does not usually change over a short-T horizon. However, 20% of the sample considered in Angrist and Newey (1991) experiences changes in schooling.

We use a revised version of the sample used in Angrist and Newey (1991) with 1087 young men. The descriptive statistics are reported in Table 5 and are quite similar to their counterparts reported in Angrist and Newey (1991, Table 1). The main ANACOVA specification of Mincer's human capital earnings function in Angrist and Newey (1991) is given by the following

$$Y_{it} = \lambda_t + \beta_1 S_{it} + \beta_2 S_{it}^2 + \beta_3 Age_{it}^2 + \beta_4 S_{it} * Age_{it} + \beta_5 U_{it} + \mathcal{A}_i + \mathcal{U}_{it} \tag{25}$$

where Y is log earnings and S is years of completed education (highest grade completed) and U is a dummy variable for union status. The above model is nonlinear in schooling, however it is linear in the unobservables, $\mathcal{A}_i$ and $\mathcal{U}_{it}$. We replicate the ANACOVA estimates of the above equation and compute the over-identification test proposed by Angrist and Newey (1991) in Table 6.[22] Column (1) refers to the estimation of the above specification, Column (2) refers to a restricted version of it, and Column (3) is a regression that only includes S, Age^2 and U.[23] As in Angrist and Newey (1991), the over-identification test is rejected, which implies that the linear fixed effects model cannot identify returns to schooling. In the following section, we revisit this application and test the time homogeneity assumption in order to examine whether the APE of schooling is nonparametrically identified.

4.2 Testing Time Homogeneity: Returns to Schooling (NLSY, 1983-1987)

The linear specification is the most widely used specification of Mincer's equation. Card (1999) however points out that there is no economic justification for the linear specification and cites empirical findings of possible nonlinearities in the relationship between schooling and earnings. Here we consider two variants of the following model exhibiting time homo-

[22] Note that the coefficient on Age_{it} cannot be separately identified from the time effects λ_t, which explains its omission from the regression table.

[23] Our replication exercise shows that the full revised sample delivers qualitatively similar results to Angrist and Newey (1991). Quantitatively, most parameter estimates are similar, except for the coefficient on S in RF. Table 6 in the supplementary appendix shows a few other replications that deliver more similar results to Angrist and Newey (1991). For details on the reasoning behind the different specifications, see Angrist and Newey (1991).

geneity with a parallel trend, where Y_{it} is log earnings,

$$
\begin{aligned}
Y_{it} &= \xi(X_{it}, \mathcal{A}_i, \mathcal{U}_{it}) + \lambda_t, \\
\mathcal{U}_{it}|X_i, \mathcal{A}_i &\overset{d}{=} \mathcal{U}_{i1}|X_i, \mathcal{A}_i.
\end{aligned}
$$

In the first variant, we only include schooling, i.e. $X_{it} = S_{it}$. In the second variant, we include both schooling and union status, i.e. $X_{it} = (S_{it}, U_{it})'$.[24]

The model implies that stayers should have the same distribution for log earnings across time, once appropriately demeaned. Formally, the testable implication is given as follows

$$
H_0 : F_{Y_{i,t-1}|\Delta X_{it}}(.|0) = F_{Y_{it}-\Delta\lambda_t|\Delta X_{it}}(.|0), \text{ for } t = 2, \ldots, T. \tag{26}
$$

For $X_{it} = S_{it}$, the stayers are individuals that do not change their schooling status. In the second variant, the stayers are individuals that neither change their schooling nor their union status.

To test the above null, Table 7 reports the bootstrap-adjusted p-values of the KS and CM statistics for the full sample (1983-1987) using the statistics in (18) and (20), where we let $X_{it} = S_{it}$ and $X_{it} = (S_{it}, U_{it})'$. For the CM statistic, we use densities $N(6.5, 0.25)$ and $N(6.5, 0.5)$. We use 200 bootstrap replications and the same grid for $\mathbb{Y}$ defined in the simulation section. For all statistics we consider for the full sample, we do not reject time homogeneity with a parallel trend at the 10% and 25% level of significance when $X_{it} = S_{it}$ and $X_{it} = (S_{it}, U_{it})'$, respectively. The Monte Carlo results in Section 3.3.2 indicate that the test statistics have good finite-sample power properties in a design similar to this empirical example. Hence, the non-rejection is unlikely to be driven by low finite-sample power of the statistics.

In addition, Table 7 reports the bootstrap-adjusted p-values for the two-period statistics, 1983-84, 1984-85, 1985-86, and 1986-87, for the KS and CM statistics given in (11) and (15). For both variants of the model, the p-values of the two-period statistics are greater than 5%. We also report the p-value of an F test for every two-period combination, which is based on

[24] Note that we cannot include *Age* in this model, since there would be no stayer subpopulations.

another implication of time homogeneity with a parallel trend,

$$E[\Delta Y_{it}|(X_{i,t-1}, X_{it}) = (x, x)] = E[\Delta Y_{it}|(X_{i,t-1}, X_{it}) = (x', x')] \;\; \forall x, x' \in \mathbb{X}, x \neq x', \tag{27}$$

The F-statistic for all two-period combinations does not reject the implication in (27) at the 5% level of significance as well.

Using the first variant of our model, we can estimate the APE of schooling for every two-period combination as follows

$$\hat{\beta}_t(\Delta S_{it} = 1) = \sum_{i=1}^{n} \Delta Y_{it} 1\{\Delta S_{it} = 1\} / \sum_{i=1}^{n} 1\{\Delta S_{it} = 1\} - \Delta \lambda_{n,t} \tag{28}$$

where $\Delta \lambda_{n,t} = \sum_{i=1}^{n} \Delta Y_{it} 1\{\Delta S_{it} = 0\} / \sum_{i=1}^{n} 1\{\Delta S_{it} = 0\}$.[25] The standard errors are computed under the assumption of cross-sectional independence.[26]

The second variant of our model allows us to estimate the APE of schooling holding union status constant. Hence, we only use individuals that increase their schooling ($\Delta S_{it} = 1$) but do not change their union status ($\Delta U_{it} = 0$) as follows

$$\begin{aligned} &\hat{\beta}_t(\Delta S_{it} = 1 | \Delta U_{it} = 0) \\ =&\sum_{i=1}^{n} \Delta Y_{it} 1\{\Delta S_{it} = 1, \Delta U_{it} = 0\} / \sum_{i=1}^{n} 1\{\Delta S_{it} = 1, \Delta U_{it} = 0\} - \Delta \lambda_{n,t}^{u} \end{aligned} \tag{29}$$

where the parallel trend is estimated using individuals that neither change their schooling or union status, specifically $\Delta \lambda_{n,t}^{u} = \sum_{i=1}^{n} \Delta Y_{it} 1\{\Delta S_{it} = \Delta U_{it} = 0\} / \sum_{i=1}^{n} 1\{\Delta S_{it} = \Delta U_{it} = 0\}$.

Table 8 reports the estimates of the APE from the two variants of our model.[27] The estimates of the APE in both cases deliver qualitatively similar results. The APE of schooling is only statistically significant in 1985-86. The APE of schooling when controlling for union

[25]Since schooling only increases by unit increments only, $\Delta S_{it} = 1$ characterizes all mover subpopulations.

[26]Since the statistics are computed for each two-period combination using first differences of individual observations, there is no need for clustering. Each APE estimate is simply a cross-sectional average of ΔY_{it}. Furthermore, $\Delta \lambda_{n,t}$ uses the stayer subpopulations, which are independent of the mover subpopulations. Hence, the two terms in $\hat{\beta}_t(\Delta S_{it} = 1)$ are independent.

[27]Estimates of the APE for different initial levels of schooling are reported in Table 7 and 8 in the supplementary appendix.

status using (29) is smaller in magnitude than the APE using (28). This is not surprising, since in the latter we average over all individuals that increased their education regardless of whether they changed their union status or not.

For illustration purposes, we also compute the APE of joining or leaving the union holding schooling constant ($\Delta S_{it} = 0$), which the second variant of our model allows us to estimate as follows, for $\Delta u \in \{-1, 1\}$.

$$\begin{aligned}
&\hat{\beta}_t(\Delta U_{it} = \Delta u | \Delta S_{it} = 0) \\
=&\sum_{i=1}^{n} \Delta Y_{it} 1\{\Delta U_{it} = \Delta u, \Delta S_{it} = 0\} / \sum_{i=1}^{n} 1\{\Delta U_{it} = \Delta u, \Delta S_{it} = 0\} - \Delta \lambda_{n,t}^{u}, \qquad (30)
\end{aligned}$$

where $\Delta u = 1$ yields an estimate of the APE of joining the union and $\Delta u = -1$ yields an estimate of the APE of leaving the union. The estimates of the APE of joining and leaving the union are reported in Table 9.[28] Except in 1985-86, we find statistically significant results for the APE of joining the union ranging between 0.102 and 0.176. As for the APE of leaving the union, we find statistically significant results in 1983-84 and 1985-86, which are -0.114 and -0.22. It is worth noting that the ANACOVA coefficient on U_{it} in Table 6 is about 0.14, which lies in the range of the nonparametric APE estimates.

5 Conclusion

This paper contributes to the literature on nonparametric identification of APEs by proposing tests of identifying assumptions, such as time homogeneity (Chernozhukov, Fernandez-Val, Hahn, and Newey, 2013) and correlated random effects (Altonji and Matzkin, 2005). The bootstrap critical values of the KS and CM statistics are shown to be asymptotically valid and perform well in finite samples. The specification tests proposed here have some special features. First, they impose minimal assumptions on the structural relationship between the outcome variable, regressors and unobservables. They also do not rely on over-identifying restrictions on the object of interest, the APE. The empirical application revisiting Angrist

[28]The estimates of the APE conditional on certain schooling levels are reported in Tables 9 and10 in the supplementary appendix.

and Newey (1991) illustrates the merits of testing nonparametric identification from an empiricist's perspective. Over-identification tests that rely on parametric assumptions have a well-known weakness. Their rejection may be due to the misspecification of the parametric model imposed on the data, even if the object of interest is nonparametrically identified. The rejection of the tests provided here has one clear interpretation; it is evidence against the identifying assumption in question.

Acknowledgments

I am grateful for the guidance and support of Graham Elliott and Andres Santos. I am especially thankful to my advisor Graham Elliott for nurturing me intellectually. I would also like to thank Prashant Bharadwaj, Julie Cullen, Bryan Graham, Ivana Komunjer, Douglas Miller, and Aaron Smith for helpful discussions, and Colin Carter, Richard Carson, Sarojini Hirshleifer, Philip Neary, Leah Nelson, Kevin Song, seminar participants at UC Berkeley, UCSD, UCSB, UC Davis, UC Riverside, NYU Abu Dhabi, and Georgetown Qatar for helpful comments. Finally, I am especially grateful to Prashant Bharadwaj for sharing the dataset used here.

A Mathematical Proofs

A.1 Proofs of Section 2 Results

Proof (Lemma 2.1) We first decompose $E[Y_{i2} - Y_{i1}|X_i = (x, x')]$ as follows

$$\begin{aligned} E[Y_{i2} - Y_{i1}|X_i = (x, x')] &= E[Y_{i2} - Y_{i2}^x|X_i = (x, x')] + E[Y_{i2}^x - Y_{i1}|X_i = (x, x')] \\ &= \underbrace{\beta_2(x \to x'|X_i = (x, x'))}_{\text{APE for Period 2 for } X_i = (x,x')} + \underbrace{E[Y_{i2}^x - Y_{i1}|X_i = (x, x')]}_{\text{Counterfactual Trend}}. \end{aligned} \tag{31}$$

Thus, the identification of the counterfactual trend is necessary and sufficient for the identification of $\beta_2(x \to x'|X_i = (x, x'))$. The former is identified iff

$$E[Y_{i2}^x - Y_{i1}|X_i = (x, x')] - E[Y_{i2} - Y_{i1}|X_i = (x, x)] = 0$$

By definition, the above is true iff

$$\int (\xi_2(x, a, u_2) - \xi_1(x, a, u_1))(dF_{\mathcal{A}_i,\mathcal{U}_{i1},\mathcal{U}_{i2}|X_i}(a, u_1, u_2|(x, x')) - dF_{\mathcal{A}_i,\mathcal{U}_{i1},\mathcal{U}_{i2}|X_i}(a, u_1, u_2|(x, x))) = 0.$$

□

Proof (Theorem 2.1) (i) The condition in Theorem 2.1 simplifies to the following

$$\begin{aligned} &\int (\xi(x, a, u_2) + \lambda_2(x) - \xi(x, a, u_1) - \lambda_1(x)) \\ &\times \left(dF_{\mathcal{A}_i,\mathcal{U}_{i1},\mathcal{U}_{i2}|X_i}(a, u_1, u_2|(x, x')) - dF_{\mathcal{A}_i,\mathcal{U}_{i1},\mathcal{U}_{i2}|X_i}(a, u_1, u_2|(x, x))\right) = 0. \end{aligned}$$

Clearly, $\lambda_2(x) - \lambda_1(x)$ cancel out from the above integrals and the above simplifies to

$$\int (\xi(x,a,u_2) - \xi(x,a,u_1)) \times \left(dF_{\mathcal{A}_i,\mathcal{U}_{i1},\mathcal{U}_{i2}|X_i}(a,u_1,u_2|(x,x')) - dF_{\mathcal{A}_i,\mathcal{U}_{i1},\mathcal{U}_{i2}|X_i}(a,u_1,u_2|(x,x))\right) = 0.$$

Now we show that $\mathcal{U}_{i1}|X_i,\mathcal{A}_i \overset{d}{=} \mathcal{U}_{i2}|X_i,\mathcal{A}_i$ implies that for any $\underline{x}$,[29]

$$\int (\xi(x,a,u_2) - \xi(x,a,u_1))dF_{\mathcal{A}_i,\mathcal{U}_{i1},\mathcal{U}_{i2}|X_i}(a,u_1,u_2|\underline{x}) = 0.$$

Note that by the time homogeneity assumption $\mathcal{A}_i,\mathcal{U}_{i1}|X_i \overset{d}{=} \mathcal{A}_i,\mathcal{U}_{i2}|X_i$, which implies the result from the following,

$$\begin{aligned}&\int \xi(x,a,u_1)dF_{\mathcal{A}_i,\mathcal{U}_{i1},\mathcal{U}_{i2}|X_i}(a,u_1,u_2|\underline{x}) = \int \xi(x,a,u_1)dF_{\mathcal{A}_i,\mathcal{U}_{i1}|X_i}(a,u_1|\underline{x})\\ &= \int \xi(x,a,u_2)dF_{\mathcal{A}_i,\mathcal{U}_{i2}|X_i}(a,u_2|\underline{x}) = \int \xi(x,a,u_2)dF_{\mathcal{A}_i,\mathcal{U}_{i1},\mathcal{U}_{i2}|X_i}(a,u_1,u_2|\underline{x})\end{aligned} \tag{32}$$

(ii) is straightforward from the following:

$$\begin{aligned}F_{Y_{i1}-\lambda_1(x)|X_i}(y|(x,x)) &= \int 1\{\xi(x,a,u_1) \leq y\}dF_{\mathcal{A}_i,\mathcal{U}_{i1}|X_i}(a,u_1|(x,x))\\ &= \int 1\{\xi(x,a,u_2) \leq y\}dF_{\mathcal{A}_i,\mathcal{U}_{i2}|X_i}(a,u_2|(x,x))\\ &= F_{Y_{i2}-\lambda_2(x)|X_i}(y|(x,x)), \quad \forall y \in \mathbb{Y}\end{aligned}$$

where the above equalities follow from noting that the conditions of the theorem imply that $F_{\mathcal{A}_i,\mathcal{U}_{i1}|X_i} = F_{\mathcal{A}_i,\mathcal{U}_{i2}|X_i}$ and that $Y_{it}^x - \lambda_t(x) = \xi(x,a,u_t)$ for $t = 1,2$. □

Proof (Theorem 2.2) (i) The result follows by plugging in the sufficient condition given in the statement of the theorem into condition (3) from Lemma 2.1.
(ii) Note that the condition in the theorem implies that $F_{\mathcal{A}_i,\mathcal{U}_{i1}|X_{i1}}(.|x) = F_{\mathcal{A}_i,\mathcal{U}_{i1}|X_i}(.|(x,x)) = F_{\mathcal{A}_i,\mathcal{U}_{i1}|X_i}(.|(x,x'))$. It follows that

$$\begin{aligned}F_{Y_{i1}|X_i}(y|(x,x)) &= \int 1\{\xi_1(x,a,u_1) \leq y\}dF_{\mathcal{A}_i,\mathcal{U}_{i1}|X_i}(a,u_1|(x,x))\\ &= \int 1\{\xi_1(x,a,u_1) \leq y\}dF_{\mathcal{A}_i,\mathcal{U}_{i1}|X_i}(a,u_1|(x,x')) = F_{Y_{i1}|X_i}(y|(x,x')) \;\; \forall y \in \mathbb{Y}.\end{aligned}$$

□

Proof (Theorem 2.3)
(i) follows by the correlated random effects assumption and $h(\underline{x}) = h(\underline{x}')$

$$\begin{aligned}\beta_t(\underline{x}_t \to \underline{x}'_t|X_i = \underline{x}') &= \int (\xi_t(\underline{x}'_t,a,u) - \xi_t(\underline{x}_t,a,u))F_{\mathcal{A}_i,\mathcal{U}_{it}|X_i}(a,u|\underline{x}')\\ &= E[Y_{it}|X_i = \underline{x}'] - \int \xi_t(\underline{x}_t,a,u))dF_{\mathcal{A}_i,\mathcal{U}_{it}|X_i}(a,u|h(\underline{x}))\\ &= E[Y_{it}|X_i = \underline{x}'] - E[Y_{it}|X_i = \underline{x}].\end{aligned}$$

[29] Note that under arbitrary individual heterogeneity, $F_{\mathcal{A}_i|X_i}(.|(x,x)) \neq F_{\mathcal{A}_i|X_i}(.|(x,x'))$, which implies that $F_{\mathcal{A}_i,\mathcal{U}_{i1},\mathcal{U}_{i2}|X_i}(.,.,.|(x,x')) \neq F_{\mathcal{A}_i,\mathcal{U}_{i1},\mathcal{U}_{i2}|X_i}(.,.,.|(x,x))$.

(ii) Since $\underline{x}_\tau = \underline{x}'_\tau$ and $h(\underline{x}) = h(\underline{x}')$, the correlated random effects assumption implies the following

$$
\begin{aligned}
F_{Y_{i\tau}|X_i}(y|\underline{x}) &= \int 1\{\xi_\tau(\underline{x}_\tau, a, u_\tau) \leq y\} dF_{\mathcal{A}_i, \mathcal{U}_{i\tau}|X_i}(a, u_\tau|\underline{x}) \\
&= \int 1\{\xi_\tau(\underline{x}'_\tau, a, u_\tau) \leq y\} dF_{\mathcal{A}_i, \mathcal{U}_{i\tau}|X_i}(a, u_\tau|\underline{x}') = F_{Y_{i\tau}|X_i}(y|\underline{x}').
\end{aligned}
$$

□

A.2 Proofs of Section 3 Results

Proof (Theorem 3.1)
We first consider the case when $T = 2$. In order to show (i) and (ii), we first have to show that the underlying empirical and bootstrap empirical processes, given below, converge to the same tight Brownian bridge. For notational convenience, let $F_t(.|\Delta X_{i2} = 0) \equiv F_{Y_{it}|\Delta X_{i2}}(.|0)$ and $F_t(.|X_i = (x^k, x^k)) \equiv F_{Y_{it}|X_i}(.|(x^k, x^k))$.

We define the empirical and bootstrap empirical processes, $\mathbb{G}_{n|\Delta X_{i2}=0}(\Delta\Lambda_n)$ and $\hat{\mathbb{G}}_{n|\Delta X_{i2}=0}(\Delta\Lambda_n)$

$$
\begin{aligned}
\mathbb{G}_{n|\Delta X_{i2}=0}(\Delta\Lambda_n) &= \sqrt{n}(F_{1,n}(.|\Delta X_{i2} = 0) - F_1(.|\Delta X_{i2} = 0)) \\
&\quad - \sqrt{n}(F_{2,n}(., \Delta\Lambda_n|\Delta X_{i2} = 0) - F_2(., \Delta\Lambda|\Delta X_{i2} = 0)) \\
\hat{\mathbb{G}}_{n|\Delta X_{i2}=0}(\Delta\Lambda_n) &= \sqrt{n}(\hat{F}_{1,n}(.|\Delta X_{i2} = 0) - F_{1,n}(.|\Delta X_{i2} = 0)) \\
&\quad - \sqrt{n}(\hat{F}_{2,n}(., \Delta\hat{\Lambda}_n|\Delta X_{i2} = 0) - F_{2,n}(., \Delta\Lambda_n|\Delta X_{i2} = 0))
\end{aligned}
$$

Now note that

$$
\begin{aligned}
F_{2,n}(., \Delta\Lambda_n|\Delta X_{i2} = 0) &= \sum_{k=1}^{K} P_n(X_{i1} = X_{i2} = x^k) F_{2,n}(. + \Delta\lambda_n(x^k)|X_{i1} = X_{i2} = x^k) \\
F_2(., \Delta\Lambda|\Delta X_{i2} = 0) &= \sum_{k=1}^{K} P(X_{i1} = X_{i2} = x^k) F_2(. + \Delta\lambda(x^k)|X_{i1} = X_{i2} = x^k) \qquad (33)
\end{aligned}
$$

Since Assumption 3.1 holds, it follows that, for $x \in \mathbb{X}$, $F_2(. + \Delta\lambda(x)|X_{i1} = X_{i2} = x)$ is Hadamard differentiable tangentially to $\mathbb{D} \times \mathbb{Y}$ by Lemma D.1 in the supplementary appendix, where $\mathbb{D} = \{g \in \mathcal{L}^\infty(\mathcal{F}) :$ g is ρ_2-uniformly continuous$\}$, where ρ_2 is the variance metric. The Hadamard derivative is given by $\phi'_{F_{2|x}(.), \Delta\lambda(x)}(g, \epsilon) = g(. + \Delta\lambda(x)) + \epsilon f_2(. + \Delta\lambda(x)|X_{i1} = X_{i2} = x)$, where $F_{t|x}(.)$ denotes $F_t(.|X_{i1} = X_{i2} = x)$ for $t = 1, 2$. Now $F_{1|x}(.) - F_{2|x}(. + \Delta\lambda(x))$ is trivially Hadamard differentiable tangentially to $\mathbb{D} \times \mathcal{L}^\infty(\mathbb{Y}) \times \mathbb{Y}$.

Let $F_{t|x,n}(.)$ be the sample analogue of $F_{t|x}(.)$. Now noting that

$$
\sqrt{n} \begin{pmatrix} F_{1|.,n}(.) - F_{1|.}(.) \\ F_{2|.,n}(.) - F_{2|.}(.) \\ \Delta\Lambda_n - \Lambda \end{pmatrix} \rightsquigarrow \begin{pmatrix} \mathbb{G}_{1|.} \\ \mathbb{G}_{2|.} \\ \mathcal{E} \end{pmatrix} \qquad (34)
$$

where $\mathbb{G}_{1|.}$ and $\mathbb{G}_{2|.}$ are each $K \times 1$ tight Brownian bridges on $\{\mathcal{L}^\infty(\mathbb{Y})\}^K$ and $\mathcal{E}$ is a K-dimensional normal random vector. We define $\mathcal{E}(x)$ as follows, $\sqrt{n}(\Delta\lambda_n(x) - \Delta\lambda(x)) \xrightarrow{p} \mathcal{E}(x)$.

By Theorem 3.9.4 in Van der Vaart and Wellner (2000), it follows that

$$
\sqrt{n}\left(F_{1|.,n}(.) - F_{2|.,n}(. + \Delta\lambda_n(.)) - (F_{1|.}(.) - F_{2|.}(. + \Delta\lambda(.)))\right) \mapsto \mathbb{G}_{1,2|.}, \qquad (35)
$$

where $\mathbb{G}_{1,2|x} = \mathbb{G}_{1|x} - \phi'_{F_{2|x}, \Delta\lambda(x)}(\mathbb{G}_{2|x}, \mathcal{E}(x))$, which is a tight Brownian bridge.

To show the weak convergence of the bootstrap empirical process given below, we have to check the conditions in Theorem 3.9.11 in Van der Vaart and Wellner (2000),

$$\sqrt{n}\left(\hat{F}_{1|.,n}(.) - \hat{F}_{2|.,n}(. + \Delta\hat{\lambda}_n(.)) - (F_{1|.,n}(.) - F_{2|.,n}(. + \Delta\lambda_n(.)))\right) \mapsto \mathbb{G}_{1,2|.}. \tag{36}$$

The conditions in Theorem 3.9.11 include (a) Hadamard-differentiability tangentially to a subspace $\mathbb{D}\times\mathcal{L}^{\infty}(\mathbb{Y})\times\mathbb{Y}$, (b) the underlying empirical processes converge to a separable limit, and (c) Condition (3.9.9), p. 378, in Van der Vaart and Wellner (2000) holds in outer probability. (a) follows from the above. Now (b) follows by (34) and tightness, since the latter implies separability. Finally, (c) is fulfilled if the conditions of Theorem 3.6.2 hold. We can consider $(\mathbb{G}_{1|.}, \mathbb{G}_{2|.}, \mathcal{E})$ as a tight Brownian bridge on $\mathcal{L}^{\infty}(\mathcal{F}) \times \mathcal{L}^{\infty}(\mathcal{F}) \times \mathbb{Y}$, where $\mathcal{F} = \{1\{y \leq t\} : t \in \mathbb{Y}\}$. Note that $\mathcal{E}$ is finite-dimensional, it suffices to show that Theorem 3.6.2 applies to $\mathcal{F}$. Since $\mathcal{F}$ is clearly Donsker and $\sup_{f\in\mathcal{F}} |\int (f - \int f dF_{t|x}(y))^2 dF_{t|x}(y)| < \infty$, the conditions in Theorem 3.6.2. hold. Thus, (c) holds, which implies (36).

Now we relate (35) and (36) to $\mathbb{G}_{n|\Delta X_{i2}=0}$ and $\hat{\mathbb{G}}_{n|\Delta X_{i2}=0}$, respectively.

$$\begin{aligned}
&\mathbb{G}_{n|\Delta X_{i2}=0}\\
&= \sqrt{n}(F_{1,n}(.|\Delta X_{i2}) - F_{2,n}(., \Delta\Lambda_n|\Delta X_{i2} = 0)) - \sqrt{n}(F_1(.|\Delta X_{i2} = 0) - F_2(., \Delta\Lambda|\Delta X_{i2} = 0))\\
&= \sqrt{n}\sum_{k=1}^{K} P_n(X_{i1} = X_{i2} = x^k|\Delta X_{i2} = 0)\left(F_{1|x^k,n}(.) - F_{2|x^k,n}(. + \Delta\lambda_n(x^k))\right)\\
&- \sqrt{n}\sum_{k=1}^{K} P_n(X_{i1} = X_{i2} = x^k|\Delta X_{i2} = 0)\left(F_{1|x^k}(.) - F_{2|x^k}(. + \Delta\lambda(x^k))\right)
\end{aligned}$$

$$\begin{aligned}
&= \sqrt{n}\sum_{k=1}^{K} P(X_{i1} = X_{i2} = x^k|\Delta X_{i2} = 0)\left(F_{1|x^k,n}(.) - F_{2|x^k,n}(. + \Delta\lambda_n(x^k))\right)\\
&- \sqrt{n}\sum_{k=1}^{K} P(X_{i1} = X_{i2} = x^k|\Delta X_{i2} = 0)\left(F_{1|x^k}(.) - F_{2|x^k}(. + \Delta\lambda(x^k))\right)\\
&+ \sum_{k=1}^{K}(P_n(X_{i1} = X_{i2} = x^k|\Delta X_{i2} = 0) - P(X_{i1} = X_{i2} = x^k|\Delta X_{i2} = 0))\\
&\times \sqrt{n}\left(F_{1|x^k,n}(.) - F_{2|x^k,n}(. + \Delta\lambda_n(x^k)) - (F_{1|x^k}(.) - F_{2|x^k}(. + \Delta\lambda(x^k)))\right) \qquad (37)
\end{aligned}$$

$$\begin{aligned}
&\rightsquigarrow \sum_{k=1}^{K} P(X_{i1} = X_{i2} = x^k|\Delta X_{i2} = 0)(\mathbb{G}_{1|x^k} - \phi'_{F_2,\Delta\lambda(x^k)}(\mathbb{G}_{2|x^k}, \mathcal{E}(x^k)))\\
&\equiv \mathcal{H}(\Delta\Lambda) \qquad (38)
\end{aligned}$$

since the last term of the last equality converges to zero in probability by the weak convergence of (37) and $P_n(X_{i1} = X_{i2} = x^k|\Delta X_{i2} = 0) \xrightarrow{p} P(X_{i1} = X_{i2} = x^k|\Delta X_{i2} = 0)$ for all $k = 1, 2, ..., K$. Hence, we have shown that the empirical process, $\mathbb{G}_{n|\Delta X_{i2}=0}$ converges to a tight Brownian bridge.

Now the bootstrap empirical process can be decomposed as follows

$$\begin{aligned}
&\hat{\mathbb{G}}_{n|\Delta X_{i2}=0}\\
&= \sqrt{n}\left(\hat{F}_{1,n}(.|\Delta X_{i2} = 0) - \hat{F}_{2,n}(., \Delta\hat{\Lambda}_n|\Delta X_{i2} = 0) - \{F_{1,n}(.|\Delta X_{i2} = 0) - F_{2,n}(., \Delta\Lambda_n|\Delta X_{i2} = 0)\}\right)\\
&= \sqrt{n}\sum_{k=1}^{K} \hat{P}_n(X_{i1} = X_{i2} = x^k|\Delta X_{i2} = 0)\\
&\times \left(\hat{F}_{1|x^k,n}(. + \Delta\hat{\lambda}_n(x^k)) - \hat{F}_{2|x^k,n}(.) - (F_{1|x^k,n}(. + \Delta\lambda_n(x^k)) - F_{2|x^k,n}(.))\right)
\end{aligned}$$

$$
\begin{aligned}
&= \sqrt{n}\sum_{k=1}^{K} P(X_{i1}=X_{i2}=x^k|\Delta X_{i2}=0)\\
&\times \left(\hat{F}_{1|x^k,n}(.)-\hat{F}_{2|x^k,n}(.+\Delta\hat{\lambda}_n(x^k))-(F_{1|x^k,n}(.)-F_{2|x^k,n}(.+\Delta\lambda_n(x^k)))\right)\\
&+\sum_{k=1}^{K}(\hat{P}_n(X_{i1}=X_{i2}=x^k|\Delta X_{i2}=0)-P(X_{i1}=X_{i2}=x^k|\Delta X_{i2}=0))\\
&\times \sqrt{n}\left(\hat{F}_{1|x^k,n}(.)-\hat{F}_{2|x^k,n}(.+\Delta\hat{\lambda}_n(x^k))-(F_{1|x^k,n}(.)-F_{2|x^k,n}(.+\Delta\lambda_n(x^k)))\right)\\
&\rightsquigarrow \mathcal{H}(\Delta\Lambda)
\end{aligned}
\tag{39}
$$

where the first term of the last equality follows by the continuous mapping theorem and (36). The second term converges to zero by (36) and $(\hat{P}_n(X_{i1}=X_{i2}=x^k|\Delta X_{i2}=0)-P(X_{i1}=X_{i2}=x^k|\Delta X_{i2}=0))\xrightarrow{p}0$ by Lemma D.2 in Section D.3 in the supplementary appendix. Thus, the bootstrap empirical process converges to the same tight Brownian bridge as the empirical process.

Since $\|.\|_{\infty,\mathbb{Y}}$ and $\|.\|_{2,\phi}$ are continuous, convex functionals, it follows that $\|\mathcal{H}\|_{\infty,\mathbb{Y}}$ and $\|\mathcal{H}\|_{2,\phi}$ have absolutely continuous and strictly increasing distributions on their support $[0,\infty)$, except possibly at zero, by Theorem 11.1 in Davydov, Lifshits, and Smorodina (1998). Since $F_1(.|\Delta X_{i2}=0)$ and $F_2(.|\Delta X_{i2}=0)$ are non-degenerate, then $P(\|\mathcal{H}\|_{\infty,\mathbb{Y}}=0)=0$ and $P(\|\mathcal{H}\|_{2,\phi}=0)=0$. Thus, both norms of $\mathcal{H}$ have absolutely continuous distributions on $[0,\infty)$. Now the critical values of the bootstrap-adjusted KS and CM tests are given by

$$
\begin{aligned}
\hat{c}_n^{KS} &= \inf\{t:\hat{P}_n(\|\hat{\mathbb{G}}_{n|\Delta X_{i2}=0}\|_{\infty,\mathbb{Y}}>t)\le\alpha\},\\
\hat{c}_n^{CM} &= \inf\{t:\hat{P}_n(\|\hat{\mathbb{G}}_{n|\Delta X_{i2}=0}\|_{2,\phi}>t)\le\alpha\},
\end{aligned}
$$

where $\hat{P}_n$ is the bootstrap probability measure for the sample. Given the above, it follows that

$$
\begin{aligned}
\hat{c}_n^{KS} &\xrightarrow{p} c^{KS} = \inf\{t:P(\|\mathcal{H}(\Delta\Lambda)\|_{\infty,\mathbb{Y}}>t)\le\alpha\},\\
\hat{c}_n^{CM} &\xrightarrow{p} c^{CM} = \inf\{t:P(\|\mathcal{H}(\Delta\Lambda)\|_{2,\phi}>t)\le\alpha\}.
\end{aligned}
$$

Thus, under the null, the bootstrap-adjusted critical values yield correct asymptotic size. Hence, we have shown (i). By the tightness of the limit process, it follows that $\hat{c}_n^{KS}$ and $\hat{c}_n^{CM}$ are bounded in probability. The statistics $KS_{n,\mathbb{Y}}(\Delta\Lambda_n)$ and $CM_{n,\phi}(\Delta\Lambda_n)$ clearly diverge to infinity under an alternative hypothesis. Thus, the tests are consistent against any fixed alternative, which proves (ii).

The extension to the case where $T>2$ is straightforward since the statistic in (7) are convex functions of two-period statistics for $(t-1,t)$, where $t=2,\ldots T$. Hence, Theorem 11.1 Davydov, Lifshits, and Smorodina (1998) applies, which implies that the statistics in (7) for $T>2$ have absolutely continuous and strictly increasing distributions. Hence the bootstrap-adjusted critical values yield correct asymptotic size and are consistent against fixed alternatives by similar arguments to the $T=2$ case. □

Proof (Theorem 3.2)
We first have to show that the underlying empirical and bootstrap empirical processes converge to the same tight Brownian bridge. Let $m_{k,l}\equiv|\{\underline{x}\in\mathbb{X}^T_{k,l,t}:\underline{x}\in\mathbb{X}^T\}|$. Our statistics can be written as follows:

$$
KS_{n,\mathbb{Y}}=\frac{1}{KT}\sum_{t=1}^{T}\sum_{k=1}^{K}\sum_{l=1}^{L}P_n(h(X_i)=h_l)\sum_{j=1}^{m_{k,l}}P_n(X_i=\underline{x}_j|h(X_i)=h_l)\|\sqrt{n}\{F_{t,n}(.|\underline{x}_j)-\bar{F}_{t,n,l}(.)\}\|_{\infty,\mathbb{Y}}
$$

$$CM_{n,\phi} = \frac{1}{KT}\sum_{t=1}^{T}\sum_{k=1}^{K}\sum_{l=1}^{L} P_n(h(X_i)=h_l)\sum_{j=1}^{m_{k,l}} P_n(X_i=\underline{x}_j|h(X_i)=h_l)\|\sqrt{n}\{F_{t,n}(.|\underline{x}_j)-\bar{F}_{t,n,l}(.)\}\|_{2,\phi}.$$

Let $(\zeta(.,\underline{x}))_{\underline{x}\in\mathbb{X}^T}$ be the vector that contains the elements $\{\zeta(.,\underline{x}):\underline{x}\in\mathbb{X}^T\}$.

$$\begin{pmatrix} \sqrt{n}\left(F_{1,n}(.|X_i=\underline{x})-F_1(.|X_i=\underline{x})\right)_{\underline{x}\in\mathbb{X}^T} \\ \sqrt{n}\left(F_{2,n}(.|X_i=\underline{x})-F_2(.|X_i=\underline{x})\right)_{\underline{x}\in\mathbb{X}^T} \\ \dots \\ \sqrt{n}\left(F_{T,n}(.|X_i=\underline{x})-F_T(.|X_i=\underline{x})\right)_{\underline{x}\in\mathbb{X}^T} \end{pmatrix} \rightsquigarrow \mathbb{G} \tag{40}$$

Since $T<\infty$ and $|\mathbb{X}^T|<\infty$, the joint distribution of the centered empirical conditional cdfs converges to a tight Brownian bridge. Now note that a linear combination of the above yields the empirical process that we use to construct our statistics.

$$\left(\left(\begin{pmatrix} \sqrt{n}(F_{1,n}(.|\underline{x}_j)-\bar{F}_{1,n,l}(.)-(F_1(.|\underline{x}_j)-\bar{F}_{1,l}(.))) \\ \dots \\ \sqrt{n}(F_{T,n}(.|\underline{x}_j)-\bar{F}_{T,n,l}(.)-(F_T(.|\underline{x}_j)-\bar{F}_{T,l}(.))) \end{pmatrix}_{j\in\{1,2,\dots,m_{k,l}\}}\right)_{l\in\{1,2,\dots,L\}}\right)_{k\in\{1,\dots,K\}}$$
$$\rightsquigarrow \mathcal{H},$$

where $\mathcal{H}\equiv((\mathcal{H}_{j,l})_{j=1,..,m_{k,l}})_{l=1,..,L}$. Note that the above process is a $(T\sum_{k=1}^{K}\sum_{l=1}^{L}m_{k,l})\times 1$ vector of functionals. Since all of the above processes are defined on a Donsker class, the bootstrap empirical process also converges to the same limit process by Theorem 3.6.1 in Van der Vaart and Wellner (2000).

$$\left(\left(\begin{pmatrix} \sqrt{n}(\hat{F}_{1,n}(.|\underline{x}_j)-\hat{\bar{F}}_{1,n,l}(.)-(F_{1,n}(.|\underline{x}_j)-\bar{F}_{1,n,l}(.))) \\ \dots \\ \sqrt{n}(\hat{F}_{T,n}(.|\underline{x}_j)-\hat{\bar{F}}_{T,n,l}(.)-(F_{T,n}(.|\underline{x}_j)-\bar{F}_{T,n,l}(.))) \end{pmatrix}_{j\in\{1,2,\dots,m_{k,l}\}}\right)_{l\in\{1,2,\dots,L\}}\right)_{k\in\{1,\dots,K\}}$$
$$\rightsquigarrow \mathcal{H}.$$

Now we give the limiting statistics as follows,

$$KS_{\mathbb{Y}} = \frac{1}{KT}\sum_{t=1}^{T}\sum_{k=1}^{K}\sum_{l=1}^{L} P(h(X_i)=h_l)\sum_{j=1}^{m_{k,l}} P(X_i=\underline{x}_j|h(X_i)=h_l)\|\mathcal{H}_{j,l}\|_{\infty,\mathbb{Y}},$$

$$CM_{\phi} = \frac{1}{KT}\sum_{t=1}^{T}\sum_{k=1}^{K}\sum_{l=1}^{L} P(h(X_i)=h_l)\sum_{j=1}^{m_{k,l}} P_n(X_i=\underline{x}_j|h(X_i)=h_l)\|\mathcal{H}_{j,l}\|_{2,\phi}.$$

Since the above is a linear combination of convex continuous functionals, it follows that Theorem 11.1 in Davydov, Lifshits, and Smorodina (1998) applies. Thus, the distributions of $KS_{\mathbb{Y}}$ and CM_{ϕ} are absolutely continuous and strictly increasing on $(0,\infty)$. Since $F_t(.|X_i=\underline{x})$ is non-degenerate for $\underline{x}\in\mathbb{X}^T$ and $t=1,2...,T$, it follows that $P(KS_{\mathbb{Y}}=0)=0$ and $P(CM_{\phi}=0)=0$. Hence, it follows that their distribution is absolutely continuous on $[0,\infty)$. Now it remains to show that $KS_{n,\mathbb{Y}}$ and $CM_{n,\phi}$ converge to $KS_{\mathbb{Y}}$ and CM_{ϕ}, respectively.
Let T_n with norm $\|.\|$ denote either the KS or CM with their respective norms, and let $\mathcal{H}_{n,j,l}$ denote

the relevant empirical process

$$\begin{aligned}
T_n &= \frac{1}{KT}\sum_{t=1}^{T}\sum_{k=1}^{K}\sum_{l=1}^{L} P_n(h(X_i)=h_l)\sum_{j=1}^{m_{k,l}} P_n(X_i=\underline{x}_j|h(X_i)=h_l)\|\mathcal{H}_{n,j,l}\| \\
&= \frac{1}{KT}\sum_{t=1}^{T}\sum_{k=1}^{K}\sum_{l=1}^{L} P(h(X_i)=h_l)\sum_{j=1}^{m_{k,l}} P(X_i=\underline{x}_j|h(X_i)=h_l)\|\mathcal{H}_{n,j,l}\| \\
&+ \frac{1}{KT}\sum_{t=1}^{T}\sum_{k=1}^{K}\sum_{l=1}^{L} P_n(h(X_i)=h_l)\sum_{j=1}^{m_{k,l}} (P_n(X_i=\underline{x}_j|h(X_i)=h_l)-P(X_i=\underline{x}_j|h(X_i)=h_l))\|\mathcal{H}_{n,j,l}\| \\
&+ \frac{1}{KT}\sum_{t=1}^{T}\sum_{l=1}^{L} (P_n(h(X_i)=h_l)-P(h(X_i)=h_l))\sum_{j=1}^{m_{k,l}} P_n(X_i=\underline{x}_j|h(X_i)=h_l)\|\mathcal{H}_{n,j,l}\| \\
&\rightsquigarrow \mathcal{T}
\end{aligned}$$

where $\mathcal{T}$ equals $KS_{\mathbb{Y}}$ and CM_ϕ for the KS and CM statistics, respectively. The convergence follows since the latter two terms converge in probability to zero, since $(P_n(X_i = \underline{x}_j|h(X_i) = h_l) - P(X_i = \underline{x}_j|h(X_i) = h_l)) \xrightarrow{p} 0$ and $(P_n(h(X_i) = h_l) - P(h(X_i) = h_l)) \xrightarrow{p} 0$, and both terms are multiplied by $O_p(1)$ terms. (i) and (ii) follow by similar arguments as in Theorem 3.1. □

References

Abadie, A. (2002): "Bootstrap Tests for Distributional Treatment Effects in Instrumental Variable Models," *Journal of the American Statistical Association*, 97(457), pp. 284–292.

Altonji, J., and R. Matzkin (2005): "Cross-section and Panel Data Estimators for Nonseparable Models with Endogenous Regrssors," *Econometrica*, 73(3), 1053–1102.

Andrews, D. W. K. (1997): "A Conditional Kolmogorov Test," *Econometrica*, 65(5), pp. 1097–1128.

Angrist, J. D. (2001): "Estimation of Limited Dependent Variable Models with Dummy Endogenous Regressors: Simple Strategies for Empirical Practice," *Journal of Business and Economic Statistics*, 19(1), 2–16.

Angrist, J. D., and W. K. Newey (1991): "Over-identification Tests in Earnings Functions with Fixed Effects," *Journal of Business and Economic Statistics*, 9(3), 317–323.

Arellano, M., and S. Bonhomme (2011): "Nonlinear Panel Data Analysis," *Annual Review of Economics*, 3, 395–424.

Arellano, M., and B. Honore (2001): "Panel Data Models: Some Recent Developments," in *Handbook of Econometrics*, ed. by J. Heckman, and E. Leamer, vol. 5. Elsevier Science.

Athey, S., and G. Imbens (2006): "Identification and Inference of Nonlinear Difference-in-Difference Models," *Econometrica*, 74(2), 431–497.

Bester, C. A., and C. Hansen (2009): "Identification of Marginal Effects in a Nonparametric Correlated Random Effects Model," *Journal of Business and Economic Statistics*, 27(2), 235–250.

Bonhomme, S. (2012): "Functional Differencing," *Econometrica*, 80(4), 1337–1385.

Card, D. (1999): "The causal effect of education on earnings," *Handbook of Labor Economics*.

Chamberlain, G. (1982): "Multivariate regression models for panel data," *Journal of Econometrics*, 18(1), 5–46.

——— (1984): "Panel data," in *Handbook of Econometrics*, ed. by Z. Griliches, and M. Intriligator, pp. 1247–1318. Elsevier.

——— (2010): "Binary Response Models for Panel Data: Identification and Information," *Econometrica*, 78(1), 159–168.

Chernozhukov, V., I. Fernandez-Val, J. Hahn, and W. Newey (2013): "Average and Quantile Effects in Nonseparable Panel Data Models," *Econometrica*, 81(2), pp.535–580.

Chernozhukov, V., I. Fernandez-Val, S. Hoderlein, S. Holzmann, and W. Newey (2015): "Nonparametric Identification in Panels using Quantiles," *Journal of Econometrics.*

Chernozhukov, V., H. Hong, and E. Tamer (2007): "Estimation and Confidence Regions for Parameter Sets in Econometric Models," *Econometrica*, 75(5), 1243–1284.

Davydov, Y., M. Lifshits, and N. Smorodina (1998): *Local Properties of Distributions of Stochastic Functionals.* American Mathematical Society, Providence.

d'Haultfoeuille, X., S. Hoderlein, and Y. Sasaki (2013): "Nonlinear Difference-in-Differences in Repeated Cross Sections with Continuous Treatment," *Cemmap Working Paper.*

Evdokimov, K. (2010): "Identification and Estimation of a Nonparametric Panel Data Model with Unobserved Heterogeneity," *Department of Economics, Princeton University.*

——— (2011): "Nonparametric Identification of a Nonlinear Panel Model with Application to Duration Analysis with Multiple Spells," *Department of Economics, Princeton University.*

Graham, B., and J. Powell (2012): "Identification and estimation of average partial effects in 'irregular' correlated random coefficient panel data models," *Econometrica*, 80(5), 2105–2152.

Hoderlein, S., and E. Mammen (2009): "Identification and estimation of local average derivatives in non-separable models without monotonicity," *Econometrics Journal*, 12(1), 1–25.

Hoderlein, S., L. Su, and H. White (2013): "Testing for Monotonicity in Unobservables under Unconfoundedness," *BC Working Paper*.

Hoderlein, S., and H. White (2012): "Nonparametric Identification of Nonseparable Panel Data Models with Generalized Fixed Effects," *Journal of Econometrics*, 168(2), 300–314.

Honore, B., and E. Kyriazidou (2000a): "Estimation of Tobit-type models with individual sepcific effects," *Econometric Reviews*, 19, 341–366.

Honore, B. E. (1992): "Trimmed Lad and Least Squares Estimation of Truncated and Censored Regression Models with Fixed Effects," *Econometrica*, 60(3), pp. 533–565.

Honore, B. E. (1993): "Orthogonality conditions for Tobit models with fixed effects and lagged dependent variables," *Journal of Econometrics*, 59(12), 35 – 61.

Honore, B. E., and E. Kyriazidou (2000b): "Panel Data Discrete Choice Models with Lagged Dependent Variables," *Econometrica*, 68(4), 839–874.

Honore, B. E., and E. Tamer (2006): "Bounds on Parameters in Panel Dynamic Discrete Choice Models," *Econometrica*, 74(3), 611–629.

Kyriazidou, E. (1997): "Estimation of a Panel Data Sample Selection Model," *Econometrica*, 65(6), pp. 1335–1364.

Lu, X., and H. White (2014): "Testing for separability in structural equations," *Journal of Econometrics*, 182(1), 14–26.

Magnac, T. (2004): "Panel Binary Variables and Sufficiency: Generalizing Conditional Logit," *Econometrica*, 72(6), 1859–1876.

Manski, C. F. (1987): "Semiparametric Analysis of Random Effects Linear Models from Binary Panel Data," *Econometrica*, 55(2), pp. 357–362.

Mundlak, Y. (1978): "On the Pooling of Time Series and Cross Section Data," *Econometrica*.

Olley, G. S., and A. Pakes (1996): "The Dynamics of Productivity in the Telecommunications Equipment Industry," *Econometrica*, 64(6), 1263–97.

Quessy, J.-F., and F. Ethier (2012): "Cramer-von Mises and characteristic fuunction tests for the two and k-samlpe problems with dependent data," *Computational Statistics and Data Analysis*, 56(6), 2097–2111.

Romano, J. P., A. M. Shaikh, and M. Wolf (2008): "Formalized Data Snooping Based on Generalized Error Rates," *Econometric Theory*, 24(2), 404–447.

Romano, J. P., and M. Wolf (2005): "Exact and Approximate Stepdown Methods for Multiple Hypothesis Testing," *Journal of the American Statistical Association*, 100(469), 94–108.

Su, L., S. Hoderlein, and H. White (2013): "Testing Monotonicity in Unobservables with Panel Data," *BC Working Paper*.

Su, L., Y. Tu, and A. Ullah (2015): "Testing Additive Separability of Error Term in Nonparametric Structural Models," *Econometric Reviews*, 34(6-10), 1057–1088.

Van der Vaart, A., and J. Wellner (2000): *Weak Convergence and Empirical Processes*. Springer-Verlag, New York.

Table 2: Baseline Simulation Results: $T = 2$, $K = 2$.

	n=500						n=2000					
	KS			CM			KS			CM		
α	0.025	0.05	0.10	0.025	0.05	0.10	0.025	0.05	0.10	0.025	0.05	0.10
						Model A						
nt	0.035	0.056	0.110	0.032	0.049	0.114	0.033	0.057	0.101	0.029	0.054	0.108
pt	0.013	0.024	0.052	0.022	0.054	0.093	0.021	0.035	0.070	0.032	0.060	0.112
gpt	0.003	0.011	0.030	0.011	0.031	0.062	0.007	0.025	0.050	0.033	0.057	0.095
cre	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
						Model B						
nt	0.366	0.492	0.630	0.360	0.467	0.590	0.933	0.971	0.989	0.945	0.971	0.993
pt	0.014	0.020	0.053	0.019	0.042	0.101	0.024	0.032	0.069	0.034	0.059	0.111
gpt	0.003	0.007	0.029	0.018	0.030	0.068	0.007	0.023	0.050	0.029	0.057	0.098
cre	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
						Model C						
nt	0.308	0.440	0.598	0.264	0.377	0.511	0.929	0.968	0.988	0.926	0.959	0.982
pt	0.342	0.450	0.584	0.344	0.452	0.567	0.970	0.986	0.993	0.967	0.986	0.993
gpt	0.003	0.008	0.029	0.017	0.033	0.067	0.009	0.026	0.049	0.032	0.054	0.099
cre	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
						Model D						
nt	0.985	0.995	1.000	0.925	0.961	0.983	1.000	1.000	1.000	1.000	1.000	1.000
pt	0.409	0.531	0.666	0.107	0.165	0.273	0.989	0.993	0.996	0.555	0.692	0.800
gpt	0.014	0.032	0.059	0.075	0.130	0.195	0.108	0.169	0.279	0.320	0.428	0.549
cre	0.035	0.063	0.114	0.025	0.044	0.106	0.029	0.052	0.109	0.027	0.043	0.092

Notes: The table reports the rejection probabilities across $1,000$ simulations using the bootstrap critical values for the statistics defined in Equations (10)-(17), where *nt*, *pt*, *gpt*, and *cre* follow the convention of the superscripts in the definitions. Bold font indicates that the model considered satisfies the null hypothesis for the statistic in question. Models A-D are defined in Table 1. The CM statistic is implemented using ϕ as the standard normal density.

Table 3: Monte Carlo Design Resembling NLSY: Means and Standard Deviations

t	1	2	3	4	5
$P(X_{i,t-1} = X_{it})$		0.84	0.93	0.91	0.96
Model B ($\lambda_{t+1} - \lambda_t = 0.05$)					
$E[Y_{it}]$	6.30	6.38	6.44	6.48	6.53
$std(Y_{it})$	(0.50)	(0.52)	(0.53)	(0.52)	(0.52)
Model B ($\lambda_{t+1} - \lambda_t = 0.1$)					
$E[Y_{it}]$	6.30	6.43	6.54	6.63	6.73
$std(Y_{it})$	(0.50)	(0.52)	(0.53)	(0.52)	(0.52)
Model C ($\lambda_{t+1} - \lambda_t = 0.05$)					
$E[Y_{it}]$	6.30	6.33	6.35	6.34	6.34
$std(Y_{it})$	(0.50)	(0.52)	(0.53)	(0.52)	(0.52)
Model C ($\lambda_{t+1} - \lambda_t = 0.1$)					
$E[Y_{it}]$	6.30	6.33	6.35	6.35	6.36
$std(Y_{it})$	(0.50)	(0.52)	(0.53)	(0.52)	(0.52)
Model D ($\lambda_{t+1} - \lambda_t = 0.05$, $\lvert\sigma_{t+1} - \sigma_t\rvert = 0.1$)					
$E[Y_{it}]$	6.30	6.39	6.43	6.46	6.53
$std(Y_{it})$	(0.50)	(0.57)	(0.52)	(0.47)	(0.52)
Model D ($\lambda_{t+1} - \lambda_t = 0.1$, $\lvert\sigma_{t+1} - \sigma_t\rvert = 0.1$)					
$E[Y_{it}]$	6.30	6.44	6.53	6.61	6.73
$std(Y_{it})$	(0.50)	(0.57)	(0.52)	(0.47)	(0.52)

Notes: All quantities in the above table are numerically calculated using a sample with $n = 10,000,000$. Models B-D are defined in Table 1 and adjusted as described in Section 3.3.2. $std(.)$ denotes the standard deviation of the variable in the brackets.

Table 4: Simulation Resembling Subsample of NLSY 1983-1987: $n = 1000$, $T = 5$, $K = 16$

	KS			CM(N(6.5,0.25))			CM(N(6.5,0.5))			CM(U(0,14))		
	0.025	0.05	0.10	0.025	0.05	0.10	0.025	0.05	0.10	0.025	0.05	0.10
						Model A						
pt	0.007	0.016	0.042	0.015	0.028	0.061	0.008	0.027	0.069	0.014	0.031	0.062
gpt	0.005	0.008	0.011	0.001	0.007	0.018	0.003	0.011	0.031	0.003	0.007	0.014
						Model B ($\lambda_{t+1} - \lambda_t = 0.005$)						
pt	0.001	0.001	0.002	0.003	0.006	0.013	0.003	0.007	0.019	0.002	0.005	0.014
gpt	0.001	0.003	0.004	0.002	0.006	0.017	0.002	0.006	0.020	0.002	0.006	0.014
						Model B ($\lambda_{t+1} - \lambda_t = 0.010$)						
pt	0.007	0.015	0.044	0.014	0.024	0.057	0.012	0.029	0.069	0.015	0.029	0.062
gpt	0.006	0.008	0.015	0.004	0.007	0.019	0.004	0.016	0.032	0.005	0.009	0.016
						Model B ($\lambda_{t+1} - \lambda_t = 0.025$)						
pt	0.007	0.019	0.042	0.013	0.030	0.063	0.015	0.028	0.064	0.016	0.032	0.066
gpt	0.005	0.008	0.016	0.004	0.006	0.017	0.004	0.012	0.027	0.003	0.006	0.015
						Model C ($\lambda_{t+1} - \lambda_t = 0.005$)						
pt	0.001	0.002	0.010	0.006	0.016	0.042	0.007	0.016	0.043	0.009	0.019	0.041
gpt	0.003	0.003	0.004	0.003	0.007	0.017	0.003	0.007	0.017	0.002	0.006	0.016
						Model C ($\lambda_{t+1} - \lambda_t = 0.010$)						
pt	0.342	0.507	0.704	0.502	0.687	0.855	0.251	0.401	0.610	0.474	0.642	0.815
gpt	0.005	0.007	0.015	0.005	0.009	0.019	0.007	0.014	0.029	0.002	0.006	0.014
						Model C ($\lambda_{t+1} - \lambda_t = 0.025$)						
pt	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
gpt	0.005	0.008	0.017	0.005	0.007	0.021	0.005	0.011	0.023	0.003	0.008	0.018
						Model D ($\lvert\sigma_{t+1} - \sigma_t\rvert = 0.025$)						
pt	0.152	0.228	0.339	0.218	0.315	0.440	0.085	0.141	0.226	0.267	0.374	0.502
gpt	0.005	0.008	0.013	0.001	0.007	0.018	0.004	0.008	0.023	0.003	0.006	0.017
						Model D ($\lvert\sigma_{t+1} - \sigma_t\rvert = 0.05$)						
pt	0.832	0.894	0.944	0.929	0.967	0.979	0.576	0.704	0.823	0.966	0.977	0.990
gpt	0.005	0.010	0.022	0.003	0.010	0.028	0.006	0.011	0.025	0.002	0.015	0.028
						Model D ($\lvert\sigma_{t+1} - \sigma_t\rvert = 0.1$)						
pt	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
gpt	0.016	0.027	0.050	0.019	0.042	0.072	0.009	0.015	0.045	0.038	0.056	0.091

Notes: The table reports rejection probabilities across 1,000 simulations using the bootstrap critical values for the statistics defined in Equations (18)-(21). Bold font indicates that the model considered satisfies the null hypothesis for the statistic in question. Models A-D are defined in Table 1. The CM statistic is implemented using the densities reported in brackets following CM in the respective column title.

Table 5: Descriptive Statistics: Returns to Schooling

	1983	1984	1985	1986	1987
Race	0.12				
Age	21.84				
	(2.22)				
S	12.34	12.45	12.57	12.57	12.61
	(1.77)	(1.83)	(1.94)	(1.94)	(1.98)
Union (U)	0.20	0.20	0.21	0.19	0.23
South	0.29	0.30	0.30	0.30	0.30
Urban	0.76	0.77	0.76	0.77	0.76
Log Hourly Wage (Y)	6.31	6.39	6.50	6.61	6.72
	(0.48)	(0.49)	(0.49)	(0.49)	(0.50)
$P_n(S_{i,t-1} = S_{it})$		0.89	0.93	0.95	0.96
$P_n(U_{it} = U_{i,t-1})$		0.85	0.86	0.88	0.85
$P_n(U_{it} = U_{i,t-1}, S_{it} = S_{i,t-1})$		0.75	0.81	0.84	0.82

Notes: The table reports cross-sectional means of each variable for 1983-1987 from the NLSY ($n = 1,087$). Standard deviations are in brackets. S is highest grade completed, South and Urban are binary variables for whether the individual lives in the South or in an urban area, respectively. $P_n(A)$ denotes the empirical probability of A.

Table 6: Returns to Schooling: ANACOVA Results

	RF (1)	X (2)	A (3)
S	-0.0688	0.0772	0.0714
	0.1308	0.0152	0.0148
S^2	-0.0070	-0.0005	
	0.0048	0.0004	
Age^2	-0.0030		-0.0001
	0.0005		0.0003
$S * Age$	0.0134		
	0.0015		
U	0.1397	0.1423	0.1424
	0.0162	0.0163	0.0163
χ^2 Statistic	150.07	149.10	176.90
(df)	(83)	(85)	(45)

Notes: This is a replication of the ANACOVA results in Angrist and Newey (1991, Table 3).

Table 7: Testing Time Homogeneity with a Parallel Trend of Log Earnings (S_{it})

Statistic ϕ	KS	CM $N(6.5, 0.25)$	CM $N(6.5, 0.5)$	F
		$X_{it} = S_{it}$		
Full Sample	0.30	0.34	0.12	
1983-84	0.49	0.16	0.08	0.11
1984-85	0.12	0.23	0.13	0.09
1985-86	0.46	0.61	0.56	0.25
1986-87	0.43	0.40	0.15	0.46
		$X_{it} = (S_{it}, U_{it})'$		
Full Sample	0.36	0.43	0.26	
1983-84	0.36	0.20	0.06	0.17
1984-85	0.08	0.11	0.08	0.05
1985-86	0.79	0.70	0.64	0.16
1986-87	0.30	0.51	0.21	0.25

Notes: The above reports the p-values of the tests for time homogeneity using the NLSY subsample, $n = 1,087$. For the KS and CM statistic, the table reports the bootstrap adjusted p-value using 200 bootstrap simulations. The p-value for the F statistic for the restriction in (27) is also reported in the last column.

Table 8: APE of Schooling on Log Hourly Wage - NLSY 1983-1987

Period	Subs	APE	S.E.	t-Stat
		$\hat{\beta}(\Delta S_{it} = 1)$		
1983-84	122	-0.012	0.043	-0.267
1984-85	73	0.095	0.055	1.723
1985-86	58	0.226	0.060	3.737
1986-87	41	-0.012	0.068	-0.179
		$\hat{\beta}(\Delta S_{it} = 1 \| U_{it} = U_{i,t-1} = 0)$		
1983-84	108	-0.001	0.040	-0.014
1984-85	61	0.075	0.051	1.468
1985-86	44	0.144	0.061	2.363
1986-87	33	0.018	0.073	0.248

Notes: The APE formulae are given in (28) and (29). *Subs* denotes subsample size.

Table 9: APE of Joining and Leaving the Union on Log Earnings - NLSY 1983-87

Period	*Subs*	APE	S.E.	t-Stat
Joining: $\hat{\beta}(\Delta U_{it} = 1 \mid \Delta S_{it} = 0)$				
1983-84	71	0.176	0.049	3.60
1984-85	67	0.132	0.068	1.96
1985-86	48	0.010	0.064	0.15
1986-87	99	0.102	0.050	2.05
Leaving: $\hat{\beta}(\Delta U_{it} = -1 \mid \Delta S_{it} = 0)$				
1983-84	73	-0.220	0.061	-3.59
1984-85	68	-0.034	0.057	-0.60
1985-86	67	-0.114	0.044	-2.56
1986-87	59	-0.071	0.060	-1.19

Notes: The formula for the APE estimates is given in (30). *Subs* denotes subsample size.