Testing Identifying Assumptions in Nonseparable Panel Data Models

Dalia A. Ghanem*

May 8, 2015

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 test statistics are bootstrap-adjusted Kolmogorov-Smirnov and Cramer-von-Mises statistics that are shown to be asymptotically valid and have good finite-sample properties in Monte Carlo simulations. An empirical application illustrates the merits of testing nonparametric identification.

JEL: C1, C14, C21, C23, C25

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

*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] The paper at hand proposes alternative specification tests that build on recent developments in the theoretical literature on nonparametric identification of average partial effects (APEs) from panel data. One of the key advantages of these tests is that they do not rely on any parametric assumptions which are likely to be misspecified in practice. Furthermore, unlike the parallel-trends assumption, the tests proposed here can be applied when only two time periods are available.

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 (Mundlak (1978), Chamberlain (1984)), 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 for the purposes of this paper. $\mathcal{A}_i$ 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 and other variables that can introduce feedback mechanisms. Equation (1) reflects the threats to identification that researchers are typically concerned about in the panel data context. The 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,

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

may change over time. Without further restrictions, we cannot identify the APE from panel data.

Fixed-effects strategies in nonseparable panel data models impose time homogeneity assumptions, which restrict the change in the structural function and the distribution of the idiosyncratic shocks across time (Hoderlein and White (2012) and Chernozhukov, Fernandez-Val, Hahn, and Newey (2013)). Correlated-random-effects strategies restrict individual heterogeneity (Altonji and Matzkin (2005) and 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 the paper at hand is to develop specification tests based on these restrictions for time homogeneity in the presence of parallel trends as well as correlated-random-effects assumptions. The testing problem extends the classical two-sample problem in statistics to the panel setting, where the two samples are dependent and the data is possible demeaned. We hence propose bootstrap-adjusted Kolmogorov-Smirnov and Cramer-von-Mises statistics and show that they are asymptotically valid.

From a practical standpoint, the choice over which identifying assumption to test falls 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 here is that subpopulations that do not experience changes in X over time (stayers) should have the same distribution of the outcome variable after correcting for the parallel trend. Observational panel data are the most widely available and 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 will 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, which falls under the correlated-random-effects category and relates to the unconfoundedness assumption. Since treatment is randomly assigned in the second time period, 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, interfered with the randomization. Hence, this test can be tailored to test for the presence of selection problems.

The specification tests proposed here have certain special features. First of all, they do not require over-identifying restrictions on the object of interest, the APE. One practical implication thereof is that some of the specification tests, such as time homogeneity and conditional-random-effects, are applicable even when $T = 2$. Secondly, the interpretation of a rejection of the specification tests proposed here is clear evidence against the identifying assumption in question. This is due to two reasons. First, the hypothesis of a given test consists of the testable restrictions of a *particular* identifying assumption. Secondly, due to the nonparametric nature and the generality 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).

Related Literature. The paper at hand builds on the recent work on fixed-T nonparametric identification of APEs, which we put here under the umbrella of "generalized first-differencing", where 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 group works such as Altonji and Matzkin (2005) and Bester and Hansen (2009), who impose correlated-random-effects assumptions, as well as Hoderlein and White (2012) and Chernozhukov, Fernandez-Val, Hahn, and Newey (2013), which impose time homogeneity. Point-identification in these papers is achieved only for a subpopulation. It is important to note that Chernozhukov, Fernandez-Val, Hahn, and Newey (2013) is 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, which include Altonji and Matzkin

(2005), Evdokimov (2010) and Evdokimov (2011).

The study of the 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.[2]

It is important to relate the differencing approach here to related 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), where the presence of a sufficient statistic for the individual effect allows for the identification of the common parameter while treating the former nonparametrically. For semiparametric binary choice models, Honore and Kyriazidou (2000b) use the intuition of generalized first-differencing to nonparametrically identify the common parameter. For the nonlinear difference-in-difference setting, Athey and Imbens (2006) exploit monotonicity assumptions on the unobservable together with time homogeneity to identify the distribution of the counterfactual from which they can derive the APE, hence their identification strategy does not fall under the generalized first-differencing approach.[3] 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.

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

[2]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).

[3]Their strategy allows them to achieve identification using time homogeneity of the idiosyncratic shocks, while allowing the structural function to vary across time.

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. The paper also builds on work using resampling methods to obtain critical values for Kolmogorov-Smirnov statistics such as Andrews (1997) and Abadie (2002).

Outline of the Paper. The rest of the paper is organized as follows. The following 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 revisiting Angrist and Newey (1991).

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 their testable restrictions. We start from the DGP in (1) and formally state the main assumptions we make in our setup. 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, it is important to note that 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. It is also worth noting that $\mathcal{A}_i$ 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 $\underline{x} \in \mathbb{X}^T$.[4] Since $|\mathbb{X}^T|$ is finite, we have finitely many subpopulations. It is important to note that each subpopulation, $\underline{x} \in \mathbb{X}^T$, is characterized by its distribution of unobservables, i.e. $F_{\mathcal{A}_i,\mathcal{U}_{i1},\mathcal{U}_{i2},...,\mathcal{U}_{iT}|X_i}(.|\underline{x})$. Hence, we can think of individuals in a subpopulation as draws from the same distribution.[5]

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 of interest here is the APE of a discrete regressor X on Y for a subpopulation $\underline{x}$. Formally, 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, $Y_{it}^x = Y_{it}(x, \mathcal{A}_i, \mathcal{U}_{it})$.

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

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,

[4] 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.

[5] It is worth mentioning that for fixed-effects strategies, one can allow $F_{\mathcal{A}_i,\mathcal{U}_{i1},\mathcal{U}_{i2},...,\mathcal{U}_{iT}|X_i}(.|\underline{x}) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} F^i_{\mathcal{A}_i,\mathcal{U}_{i1},\mathcal{U}_{i2},...,\mathcal{U}_{iT}|X_i}(.|\underline{x})$, 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.

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 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. First, we introduce the following assumption.

Assumption 2.2 *(Distribution of Unobservables)* $F_{\mathcal{A}_i,\mathcal{U}_{i1},\mathcal{U}_{i2}|X_i}$ *admits a density,* $f_{\mathcal{A}_i,\mathcal{U}_{i1},\mathcal{U}_{i2}|X_i}(a, u_1, u_2|\underline{x}) > 0 \ \forall (a, u_1, u_2, \underline{x}) \in \mathbb{A} \times \mathbb{U}^2 \times \mathbb{X}^2$.

Lemma 2.1 ***(Generalized First-Differencing)***

Let Assumptions 2.1 and 2.2 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

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

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. In order to give an intuitive explanation of (3), consider the following version of the condition which is obtained by simple manipulation and using counterfactual notation,

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

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 different subpopulations. For identification to be achieved here, these two sources of change have to "cancel each other out" on average.

The following theorem shows how time homogeneity together with a restriction on ξ_t fulfills the condition (3) and gives the testable restrictions implied by this identification approach.

Theorem 2.1 ***Fixed Effects: Identification & Testable Restrictions*** *($T = 2$)*
Let Assumptions 2.1 and 2.2 hold.

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

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 (2014), the time homogeneity assumption can be equivalently stated as $\mathcal{E}_{i1}|X_i \overset{d}{=} \mathcal{E}_{i2}|X_i$, 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 time stationarity assumptions on the structural function.[6] We will refer to $\lambda_t(x)$ as a generalized parallel trend.

[6]Athey and Imbens (2006) impose time homogeneity on the unobservables, while allowing for a nonstationary structural function. Their identification result relies on monotonicity in a scalar unobservable and inverting the distribution of the outcome variable to obtain the distribution of the unobservable. Their approach hence does not fall under the generalized first-differencing category.

The above theorem shows that time homogeneity together with the restriction on the structural function satisfy 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) will be 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, what time homogeneity together with stationarity of the structural function in unobservables ensures is that the distribution of the outcome variable does not change across time due to unobservables, which ensures that we can identify the APE from changes in the outcome variable that coincide with changes in the regressors. 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. Since the two time periods considered need not be adjacent, extension to the case where $T > 2$ is straightforward.

The testable restrictions here are not over-identifying restrictions on the object of interest, the APE. From a practical perspective, the above theorem shows that there are testable restrictions for identification using time homogeneity even when $T = 2$. The following theorem shows how a restriction on individual heterogeneity, which falls under the correlated-random-effects approach, can also identify the APE while providing testable restrictions for the two-period case. This identification approach was taken in Altonji and Matzkin (2005) and Bester and Hansen (2009).

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

Let Assumptions 2.1 and 2.2 hold.

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

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 (selection-on-observables). 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 and the treatment is randomly assigned in the second time period according to the design. 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 accommodated to that setting as in the following example.

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

Consider the following simplified example of attrition in a field experiment. Suppose that we have a simple field experiment with a single treatment, i.e. $X_{it} \in \{0, 1\}$*, and we observe the control group* $(X_i = (0, 0))$ *and treatment group* $(X_i = (0, 1))$*. The treatment is randomly assigned in the second period, hence we expect the treatment and control group to be homogeneous. 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* Z_i *be a bi-*

nary variable for whether individuals respond or not. The identification question is whether $E[Y_{i2}|X_i = (0,1), Z_i = 1] - E[Y_{i2}|X_i = (0,0), Z_i = 1]$ *can identify the APE* $\beta_2(0 \to 1|X_i = (0,1), Z_i = 1)$*. Note that by the above theorem if the conditional random effects assumption,* $\mathcal{A}_i, \mathcal{U}_{i1}, \mathcal{U}_{i2}|X_i, Z_i \stackrel{d}{=} \mathcal{A}_i, \mathcal{U}_{i1}, \mathcal{U}_{i2}|X_{i1}, Z_i$,[7] *holds, then the APE in question is identified and we have the testable restriction,* $F_{Y_{i1}|X_i,Z_i}(.|(x,x'),1) = F_{Y_{i1}|X_i,Z_i}(.|(x,x),1)$*. 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}*. The testable implication is hence that the pre-treatment outcome variable of the respondents in the control and treatment group should have the same distribution.*

The conditional-random-effects assumption is an empirically relevant example of correlated-random-effects strategies. However, there are many other possible choices of correlated-random-effects strategies. For instance, Altonji and Matzkin (2005) shows how exchangeability restrictions can be used to identify APEs. Theorem A.1 in the supplementary appendix generalizes the above theorem to the general class of correlated-random-effects assumptions in the $T > 2$ case, where $\mathcal{A}_i, \mathcal{U}_{i1}, \ldots, \mathcal{U}_{iT}|X_i \stackrel{d}{=} \mathcal{A}_i, \mathcal{U}_{i1}, \ldots, \mathcal{U}_{iT}|h(X_i)$.

3 Testing Identifying Assumptions

In this section, we develop statistics to test restrictions implied by the identifying assumptions presented in Section 2, specifically time homogeneity and correlated-random-effects assumptions. It is important to note here that the APE of a subpopulation may be identified using different assumptions. For instance, Theorems 2.1 and 2.2 show how the APE can be identified using time homogeneity and conditional-random-effects, respectively. Hence, the statistics proposed here do not test whether the APE is identified or not. They test whether a particular assumption that achieves identification holds. Hence, their rejection is clear evidence against the identifying assumption in question. As pointed out above, the empirical setting typically suggests a suitable identification strategy.

[7] It is worth pointing out here that one can omit X_{i1} from the conditioning set in this example, since $X_{i1} = 0$ for all subpopulations in this example.

The testable implications of the identifying assumptions we consider are equality restrictions on the conditional distribution of the outcome variable, hence they are as an extension of the classical two-sample problem. The Kolmogorov-Smirnov (KS) and Cramer-von-Mises (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 samples 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 statistic obtained from the b^{th} bootstrap sample.

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 the respective centered statistics for the b^{th} bootstrap sample T_n^b.*
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 here. 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.

3.1 Testing Time Homogeneity

To simplify presentation, we first focus on testing time homogeneity in the $T = 2$ case and then discuss the extension to the $T > 2$ in Section 3.1.2.

3.1.1 $T = 2$

Theorem 2.1 establishes that time homogeneity in the $T = 2$ case has the following testable restrictions

$$F_{Y_{i1}^x - \lambda_1(x)|X_i}(.|(x,x)) = F_{Y_{i2}^x - \lambda_2(x)|X_i}(.|(x,x)) \forall x \in \mathbb{X}. \tag{4}$$

Note that $|\mathbb{X}| = K$ and write $\mathbb{X} = \{x^k\}_{k=1}^K$. We can re-write the K restrictions using $\Delta\lambda(x) = \lambda_2(x) - \lambda_1(x)$ as follows

$$F_{Y_{i1}|X_i}(.|(x^k,x^k)) = F_{Y_{i2} - \Delta\lambda(x^k)|X_i}(.|(x^k,x^k)), \quad \text{for } k = 1, 2, \ldots, K. \tag{5}$$

Let $\Delta X_i = X_{i2} - X_{i1}$. We can integrate over X_i conditional on $\Delta X_i = 0$, i.e. for the stayer subpopulations. This yields a single testable restriction, which we denote by H_0^{gpt} presented below. We first need to introduce some notation. Let $\Delta\Lambda = (\Delta\lambda(x^1), ..., \Delta\lambda(x^K))'$ and

$$F_{Y_{i2}|\Delta X_i}(., \Delta\Lambda|0) = \sum_{k=1}^{K} P(X_i = (x^k, x^k)) F_{Y_{i2} - \Delta\lambda(x^k)|X_i}(.|(x^k,x^k)).$$

Now we can write H_0^{gpt} as follows, where *gpt* abbreviates the generalized parallel trend, $\lambda_t(x)$,

$$H_0^{gpt} : F_{Y_{i1}|\Delta X_i}(.|0) = F_{Y_{i2}|\Delta X_i}(., \Delta\Lambda|0). \tag{6}$$

The above restriction is an equality between the distributions of two dependent random variables that are appropriately demeaned. It is important to note here that even if the idiosyncratic shocks were independent across time, the time-invariant random variable, $\mathcal{A}_i$, remains a source of dependence between Y_{i1} and Y_{i2}.

Now to give the statistics for H_0^{gpt}, we require some additional notation. For an event A_i, $P_n(A_i) = \sum_{i=1}^n 1\{A_i\}/n$; $\Delta\lambda_n(x^k) = \frac{\sum_{i=1}^n \Delta Y_i 1\{X_i=(x^k,x^k)\}}{\sum_{i=1}^n 1\{X_i=(x^k,x^k)\}}$, $\Delta\Lambda_n = (\Delta\lambda_n(x^1), \ldots, \Delta\lambda_n(x^K))'$ and $F_{n,Y_{it}|\Delta X_i}$ is the empirical cdf of $F_{Y_{it}|\Delta X_i}$.

We denote the KS and CM statistics that test H_0^{gpt} by $KS_{n,\mathbb{Y}}^{gpt}$ and $CM_{n,\phi}^{gpt}$, respectively. Let

$$\begin{aligned} KS_{n,\mathbb{Y}}^{gpt} &= \left\|\sqrt{n}(F_{n,Y_{i1}|\Delta X_i}(.|0) - F_{n,Y_{i2}|\Delta X_i}(., \Delta\Lambda_n|0))\right\|_{\infty,\mathbb{Y}}, \\ CM_{n,\phi}^{gpt} &= \left\|\sqrt{n}(F_{n,Y_{i1}|\Delta X_i}(.|0) - F_{n,Y_{i2}|\Delta X_i}(., \Delta\Lambda_n|0))\right\|_{2,\phi}, \end{aligned} \tag{7}$$

where $\|g(y)\|_{\infty,\mathbb{Y}} = \sup_{y\in\mathbb{Y}} |g(y)|$ and $\|g(y)\|_{2,\phi} = \int_{y\in\mathbb{Y}} \int g(y)^2\phi(y)dy$. It is important to note here that both norms that define the KS and CM statistics here are non-random. In order to compute them in practice a grid for $\mathbb{Y}$ has to be chosen. For the CM statistic, ϕ is a user-chosen density. We will explore the choice of ϕ and the finite-sample performance of the CM statistic in a simulation study in Section 3.3.2).

We introduce the bootstrap centered versions of the statistics in (7). In the following, the superscript b on any sample analogue denotes that the sample analogue is computed using the b^{th} bootstrap sample. Let B denote the number of bootstrap simulations.

$$\begin{aligned} &KS_{n,\mathbb{Y}}^{gpt,b} \\ &= \|\sqrt{n}\{F_{n,Y_{i1}|\Delta X_i}^b(.|0) - F_{n,Y_{i2}|\Delta X_i}^b(., \Delta\Lambda_n^b|0) - (F_{n,Y_{i1}|\Delta X_i}(.|0) - F_{n,Y_{i2}|\Delta X_i}(., \Delta\Lambda_n|0))\}\|_{\infty,\mathbb{Y}}, \end{aligned}$$

$$CM^{gpt,b}_{n,\phi}$$
$$= \|\sqrt{n}\{F^b_{n,Y_{i1}|\Delta X_i}(.|0) - F^b_{n,Y_{i2}|\Delta X_i}(.,\Delta\Lambda^b_n|0) - (F_{n,Y_{i1}|\Delta X_i}(.|0) - F_{n,Y_{i2}|\Delta X_i}(.,\Delta\Lambda_n|0))\}\|_{2,\phi}.$$

Now it remains to show the asymptotic validity of Procedure 3.1 applied to the $KS^{gpt}_{n,\mathbb{Y}}$ and $CM^{gpt}_{n,\phi}$ using their respective bootstrap centered statistics. The key issue here is the presence of the time effects, which introduces noise into the empirical cdf. In order to ensure the convergence of the empirical process that are used to compute the KS and CM statistics, we impose the following condition. It ensures that the underlying distribution is uniformly continuous.

Assumption 3.1 *(Bounded Density)*
$F_{Y_{it}}(.)$ *has a density* $f_{Y_{it}}(.)$ *that is bounded, i.e.* $\sup_{y\in\mathbb{Y}}|f_{Y_{it}}(y)| < \infty$, $t = 1, 2$.

Theorem 3.1 *Given that* $\{Y_i, X_i\}^n_{i=1}$ *is an iid sequence,* $|\mathbb{X}| = K$, $P(\Delta X_i = 0) > 0$, $F_{Y_{it}|X_i}(.|\underline{x})$ *is non-degenerate for* $t = 1, 2$ *and* $\underline{x} \in \mathbb{X}$, *and Assumption 3.1 holds. Procedure 3.1 for* $KS^{gpt}_{n,\mathbb{Y}}$ *and* $CM^{gpt}_{n,\phi}$ *to test* H^{gpt}_0 *(i) provides correct asymptotic size* α *and (ii) is consistent against any fixed alternative.*

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-adjusted tests have correct asymptotic size and are consistent against fixed alternatives.

It is important to note that an alternative to H^{gpt}_0, which is given in (6), is to use (5) as a null hypothesis, which we refer to by $\widetilde{H}^{gpt}_0$ hereinafter. Under the assumption that $P(X_i = (x^k, x^k)) > 0$ for all $k = 1, 2, ..., K$, which holds under Assumption 2.1(iv), $\widetilde{H}^{gpt}_0$ and H^{gpt}_0 are equivalent. The statistics resulting from the former can be viewed as disaggregated versions of the statistics based on the latter. The disaggregated statistics are given by the

following,

$$\begin{aligned}
\widetilde{KS}^{gpt}_{n,\mathbb{Y}} &= \sum_{k=1}^{K} P_n(X_i = (x^k, x^k)|\Delta X_i = 0) \\
&\quad \times \left\| \sqrt{n}\{F_{n,Y_{i1}|X_i}(y|(x^k,x^k)) - F_{n,Y_{i2}-\Delta\lambda_n(x^k)|X_i}(y|(x^k,x^k))\} \right\|_{\infty,\mathbb{Y}}, \\
\widetilde{CM}^{gpt}_{n,\phi} &= \sum_{k=1}^{K} P_n(X_i = (x^k, x^k)|\Delta X_i = 0) \\
&\quad \times \left\| \sqrt{n}\{F_{n,Y_{i1}|X_i}(y|(x^k,x^k)) - F_{n,Y_{i2}-\Delta\lambda_n(x^k)|X_i}(y|(x^k,x^k))\} \right\|_{2,\phi}, \qquad (8)
\end{aligned}$$

where $F_{n,Y_{it}|X_i}$ is the empirical cdf of $F_{Y_{it}|X_i}$. The above statistics are probability-weighted averages of KS and CM statistics of the individual restrictions, respectively. It is important to note that the probability weighting is one among other weighting schemes. For instance, one could give equal weights $1/K$ to all of the individual statistics. The bootstrap procedure outlined above can also be used for $\widetilde{KS}^{gpt}_{n,\mathbb{Y}}$ and $\widetilde{CM}^{gpt}_{n,\phi}$ to adjust their respective p-values. The asymptotic validity of the bootstrap adjustment follows in a straightforward manner from the proof of Theorem 3.1. We will explore the finite-sample performance of the disaggregated statistics relative to the aggregated ones in our simulation study in Section 3.3.3.

3.1.2 $T > 2$

For $T > 2$, a DGP exhibiting time homogeneity as in Theorem 2.1 is given by

$$\begin{aligned}
Y_{it} &= \xi(X_{it}, \mathcal{A}_i, \mathcal{U}_{it}) + \lambda_t(X_{it}) \\
\mathcal{U}_{it}|X_i, \mathcal{A}_i &\overset{d}{=} \mathcal{U}_{i1}|X_i, \mathcal{A}_i, \text{for } t = 1, 2, ..., T. \qquad (9)
\end{aligned}$$

The testable implication here is given by the following

$$F_{Y_{it}|(X_{it},X_{i,t+1})}(.|(x,x)) = F_{Y_{i,t+1}-\Delta\lambda_t(x)|(X_{it},X_{i,t+1})}(.|(x,x)), \forall x \in \mathbb{X}, t = 1, 2, \ldots, T-1, \qquad (10)$$

where $\Delta\lambda_t(x) \equiv \{\lambda_{t+1}(x) - \lambda_t(x)\}$. In the above, we have $K \times (T-1)$ restrictions that we would like to test simultaneously.[8]

Let $\Delta X_{it} = X_{i,t+1} - X_{it}$. Similar to H_0^{gpt} in the $T = 2$ case, we can integrate over $(X_{it}, X_{i,t+1}) = (x, x)$ conditional on $\Delta X_{it} = 0$ to obtain the following hypothesis

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

where $\Delta\Lambda_t \equiv (\Delta\lambda_t(x^1), \Delta\lambda_t(x^2), \ldots, \Delta\lambda_t(x^K))'$. H_0^T consists of $T-1$ restrictions. Thus, the following statistics test all $T-1$ restrictions.

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

where $\Delta\Lambda_{n,t}$ is a sample analogue of $\Delta\Lambda_t$. Procedure 3.1 can be used to obtain p-values for the above statistics with the appropriate centered statistics for the bootstrap sample. Its asymptotic validity is straightforward from Theorem 3.1.

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

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

For each t, the statistics can be computed similar to the $T = 2$ case. A multiple-testing correction procedure, such as 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 $\{H_0^{gpt,t}\}_{t=1}^{T-1}$.

[8] It is important to note that (10) is a natural extension of (5) in the $T = 2$ case. It is however possible to write an more disaggregated hypothesis. Note that $F_{Y_{it}|(X_{it},X_{i,t+1})}(y|(x,x)) = \sum_{\underline{x} \in \mathbb{X}_{x,t}^T} P(X_i = \underline{x}|(X_{it}, X_{i,t+1}) = (x,x)) F_{Y_{it}|X_i}(y|\underline{x})$, where $\mathbb{X}_{x,t}^T = \{\underline{x} \in \mathbb{X}^T : (\underline{x}_t, \underline{x}_{t+1}) = (x,x), x \in \mathbb{X}, t = 1, 2, \ldots, T\}$. Thus, the testable restriction can be alternatively written as $F_{Y_{it}|X_i}(.|\underline{x}) = F_{Y_{i,t+1}-\Delta\lambda_t(x)|X_i}(.|\underline{x}),\ \forall \underline{x} \in \mathbb{X}_{x,t}^T, x \in \mathbb{X}, t = 1, 2, ..., T-1$.

3.2 Testing Correlated-Random-Effects Assumptions

Theorem 2.2 gives the testable restriction of the conditional-random-effects assumption, which falls under the correlated-random-effects category. Theorem A.1 in the Supplementary Appendix gives conditions for the identification of an APE for a subpopulation and the testable implications for correlated-random-effects assumptions. In the following, we propose test statistics for the general class of correlated-random effects assumptions. The setup for correlated-random-effects assumptions is given by the following,

$$
\begin{aligned}
Y_{it} &= \xi_t(X_{it}, \mathcal{A}_i, \mathcal{U}_{it}), \text{for } t = 1, 2, \ldots, T \\
\mathcal{A}_i, \mathcal{U}_{i1}, \mathcal{U}_{i2}, \ldots, \mathcal{U}_{iT} | X_i &\overset{d}{=} \mathcal{A}_i, \mathcal{U}_{i1}, \mathcal{U}_{i2}, \ldots, \mathcal{U}_{iT} | h(X_i). \qquad (11)
\end{aligned}
$$

where $h : \mathbb{X}^T \mapsto \mathbb{H}$. For $T = 2$ and $h(X_i) = X_{i1}$, this delivers the conditional-random-effects assumption. For identification purposes, $L \equiv |\mathbb{H}| < |\mathbb{X}^T|$, such that $\mathbb{H} = \{h_l\}_{l=1}^L$.[9]

Now we introduce some notation in order to write the testable restriction. For $l = 1, 2, \ldots, L$, 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 and have the same distribution of unobservables, since $h(\underline{x}) = h_l$. By the model given in (11), the distributions of the outcome variable at time t for all subpopulation 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.

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

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

When $|\mathbb{X}_{k,l,t}^T| > 2$, 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"

[9] Correlated-random-effects strategies impose homogeneity assumptions 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 $|\mathbb{H}| < |\mathbb{X}^T|$.

cdf. We follow their approach here and introduce the following cdf,

$$\bar{F}_{k,l,t} = \frac{1}{|\mathbb{X}^T_{k,l,t}|} \sum_{\underline{x} \in \mathbb{X}^T_{k,l,t}} F_{Y_{it}|X_i}(.|\underline{x}), \tag{12}$$

which averages over the cdfs of the outcome variable at time period t for subpopulations in the set $\mathbb{X}^T_{k,l,t}$. Recall that the subpopulations in this set have the same distribution of unobservables under the correlated-random-effects assumption in question.

The KS and CM statistics for the restrictions for time period t and x_k are given by

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

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

Averaging the above statistics over k and t, we obtain the statistics that test H^{cre}_0,

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

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

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

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

The following theorem shows that the bootstrap-adjusted test statistics 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*, the procedure described in 1-4 for* $KS^{cre}_{n,\mathbb{Y}}$ *and* $CM^{cre}_{n,\phi}$ *to test* H_0^{cre} *(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 bootstrap-adjusted 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 ($\lambda_t(x) = 0$) as well as with a parallel trend ($\lambda_t(x) = \lambda_t$ for all $x \in \mathbb{X}$).

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.[10] 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 other 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

[10] This point relates to the choice of density used to compute the CM statistic and will be discussed in greater detail in Section 3.3.2.

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$(g(x))$ $= 1\{g(x) > 0\} - 1\{g(x) < 0\}$; $\lambda_1 = 0$, $\sigma_1 = 1$.

homogeneity with a parallel trend, λ_t. Thus, Model D deviates from time homogeneity with a parallel trend 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 different choices for the density of the CM statistic and its effect on the finite-sample behavior of the statistic. Section 3.3.3 compares the aggregated and disaggregated statistics for testing time homogeneity. Finally, Section 3.3.4 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.

3.3.1 Testing Identifying Assumptions: Baseline Results

In the baseline results, we examine the finite-sample behavior of the bootstrap adjustment for $n = 500, 2000$, $T = K = 2$, and $p = 0.5$. Thus, we expect half of the sample to be stayers and half to be movers. In this design, we set $E[Y_{i1}] = 0$ by assigning μ_0 an appropriate value.[11] In terms of the models that exhibit time homogeneity, we consider Model A and two variants of each of the Models B and C, with $\lambda_2 = 0.25, 0.5$, which are about 10% and

[11] 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, $\mu_0 = 0.707$.

20% of the standard deviation of Y_{i1}, respectively.[12] We also consider two variants of Model D, with $\sigma_2 = 1.1, 1.2$. For both variants of Model D, we set $\lambda_2 = 0.5$.

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

$$KS^{nt}_{n,\mathbb{Y}} = \|F_{n,Y_{i1}|\Delta X_i}(y|0) - F_{n,Y_{i2}|\Delta X_i}(y|0)\|_{\infty,\mathbb{Y}} \tag{13}$$

$$KS^{pt}_{n,\mathbb{Y}} = \|F_{n,Y_{i1}|\Delta X_i}(y|0) - F_{n,Y_{i2}-\Delta\lambda_n|\Delta X_i}(y|0)\|_{\infty,\mathbb{Y}} \tag{14}$$

$$KS^{gpt}_{n,\mathbb{Y}} = \|F_{n,Y_{i1}|\Delta X_i}(y|0) - F_{n,Y_{i2}|\Delta X_i}(y, \Delta\Lambda_n|0)\|_{\infty,\mathbb{Y}} \tag{15}$$

$$KS^{cre}_{n,\mathbb{Y}} = \sum_{l=1}^{K} P_n(X_{i1} = x^l) \frac{1}{K} \sum_{k=1}^{K} \|F_{n,Y_{i1}|X_i}(y|(x^l, x^k)) - \bar{F}_{n,l,1}(y)\|_{\infty,\mathbb{Y}} \tag{16}$$

$$CM^{nt}_{n,\phi} = \|F_{n,Y_{i1}|\Delta X_i}(y|0) - F_{n,Y_{i2}|\Delta X_i}(y|0)\|_{2,\phi} \tag{17}$$

$$CM^{pt}_{n,\phi} = \|F_{n,Y_{i1}|\Delta X_i}(y|0) - F_{n,Y_{i2}-\Delta\lambda_n|\Delta X_i}(y|0)\|_{2,\phi} \tag{18}$$

$$CM^{gpt}_{n,\phi} = \|F_{n,Y_{i1}|\Delta X_i}(y|0) - F_{n,Y_{i2}|\Delta X_i}(y, \Delta\Lambda_n|0)\|_{2,\phi} \tag{19}$$

$$CM^{cre}_{n,\phi} = \sum_{l=1}^{K} P_n(X_{i1} = x^l) \frac{1}{K} \sum_{k=1}^{K} \|F_{n,Y_{i1}|X_i}(y|(x^l, x^k)) - \bar{F}_{n,l,1}(y)\|_{2,\phi}. \tag{20}$$

where $\bar{F}_{n,l,1}(.) = \sum_{k=1}^{K} F_{n,Y_{i1}|X_i}(.|(l, x^k))/K$, $\Delta\lambda_n = \sum_{i=1}^{n}(Y_{i2}-Y_{i1})1\{\Delta X_i = 0\}/\sum_{i=1}^{n} 1\{\Delta X_i = 0\}$. $\Delta\Lambda_n$ 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 3.3.2. 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 $\underline{y} = \min_{i,t} \tilde{Y}_{it}$ and $\bar{y} = \max_{i,t} \tilde{Y}_{it}$, where $\tilde{Y}_{it}$ denotes the appropriately demeaned Y_{it}.[13] 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

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

[12] The standard deviation of Y_{i1} is numerically calculated to be 2.6, using a sample of size $10,000,000$.
[13] For instance for the *pt* statistics, $\tilde{Y}_{i1} = Y_{i1}$ and $\tilde{Y}_{i2} = Y_{i2} - \Delta\lambda_n$.

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

Table 2 reports the rejection probabilities of the bootstrap-adjustment to the above statistics in the baseline Monte Carlo design, where $n = 500, 2000$ and $T = K = 2$. 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* for $n = 500$ tends to be under-sized.

Under Model D, the *cre* statistics provide good size control, where the CM statistic seems to fair better in this regard than the KS statistic. However, for *pt*, the KS statistic exhibits much better finite-sample power over the CM statistic, whereas the latter has better power properties over the former for *gpt*. It is important to point out here that all *gpt* statistics seem much 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 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, since $X_{it} \in \{-1,1\}$ when $K = 2$. 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 finite-sample power properties.

3.3.2 The Choice of the Density for the CM statistic

In this subsection, we examine the choice of the density in constructing the CM statistic. Here we employ the same models and choices of n and T as in the baseline design described above, while considering several different choices of the density ϕ in the CM statistic. Thus, we examine the behavior of the statistics in (17)-(20) letting ϕ be the density of the following distributions: $N(0,1)$, $N(0,3)$, $U(-2,2)$, and $U(-5,5)$. Note that in the models in our design, the mean of Y_{i1} is equal to zero, and the *pt* and *gpt* statistics considered here demean Y_{i2} to

have the same mean of Y_{i1}. This is the reasoning behind choosing densities that are centered around zero. The choice between the normal and uniform densities impacts whether the center of the distribution is weighed more heavily than the tails as in the former or whether the support of the distribution is weighed equally as in the latter. The remaining difference is in the standard deviation of the density. Here it is important to note that the standard deviation of Y_{i1} is 2.61 under Models A-C and 2.97 under Model D.[14] Hence, the choice of the $N(0,1)$ density uses a standard deviation that is smaller than the standard deviation of the underlying variables, thus emphasizing the center of the distribution to a much greater degree than the $N(0,3)$ density. Similarly, for the $U(-2,2)$ and $U(-5,5)$ densities, the former emphasizes the center of the distribution whereas the latter gives weight to a larger portion of the support of Y_{it}.

Table 3 reports the simulation results for the CM statistics using the densities mentioned above when $n = 500$. In this design, we find that increasing the $N(0,3)$ density improves over the $N(0,1)$ density in terms of power, while exhibiting good size control. The same applies for the comparison between $U(-5,5)$ and $U(-2,2)$. For the design we consider, the choice between uniform and normal densities does not seem to have any stark consequences. The power improvement from using $N(0,3)$ ($U(-5,5)$) instead of $N(0,1)$ ($U(-2,2)$) is especially large for the *pt* statistic under Model D. Examining the simulation results for the same statistics when $n = 2000$, which are reported in Table 4, we find that these favorable finite-sample power properties are accompanied by poor size control. For instance, see the rejection probabilities for *nt* and *pt* under Models A and B.

In practice, for the *nt* and *cre* statistic, the center of the density of the CM statistic has to be chosen by the user. For the *pt* and *gpt* statistics, however, this choice can be avoided if both Y_{i1} and Y_{i2} are demeaned as in the following statistics,

$$CM^{pt}_{n,\phi(0)} = \left\| \sqrt{n} \left\{ F_{n,Y_{i1}-\lambda_{n,1}|\Delta X_i}(.|0) - F_{n,Y_{i2}-\lambda_{n,2}|\Delta X_i}(.|0) \right\} \right\|_{2,\phi}, \tag{21}$$

$$CM^{gpt}_{n,\phi(0)} = \left\| \sqrt{n} \left\{ F_{n,Y_{i1}|\Delta X_i}(., \Lambda_{n,1}|0) - F_{n,Y_{i2}|\Delta X_i}(., \Lambda_{n,2}|0) \right\} \right\|_{2,\phi}, \tag{22}$$

[14] These standard deviations were numerically calculated using a sample of size $10,000,000$.

where $\lambda_{n,t} = \sum_{i=1}^{n} Y_{it} 1\{\Delta X_i = 0\} / \sum_{i=1}^{n} 1\{\Delta X_i = 0\}$ and $\Lambda_{n,t} = (\sum_{i=1}^{n} Y_{it} 1\{X_i = (x^1, x^1)\} / \sum_{i=1}^{n} 1\{X_i = (x^1, x^1)\}, \ldots, \sum_{i=1}^{n} Y_{it} 1\{X_i = (x^K, x^K)\} / \sum_{i=1}^{n} 1\{X_i = (x^K, x^K)\}$. Tables 5 and 6 report the rejection probabilities of the bootstrap adjustment of the above statistics in our design for $n = 500$ and $n = 2000$, respectively. For $n = 500$, the above *pt* statistic tends to be quite under-sized relative to the statistic in (18), which translates into poor finite-sample power, for instance see Table **??** under Model C ($\lambda_2 = 0.25$) with ϕ as the standard normal. For this setting, using a greater standard deviation such as $N(0, 3)$ improves finite-sample power without distorting size. For $n = 2000$, the above test statistics maintain good size control and finite-sample power relative to the statistics in (18) and (19).

3.3.3 Testing Time Homogeneity: Aggregated vs. Disaggregated Statistics

So far we have considered the test statistics that aggregate across all stayer subpopulations a given in (13)-(15) and (17)-(19), which we will refer to as the aggregated statistics hereinafter. As pointed out in Section 3.1, one could alternatively use the following statistics,

$$\widetilde{KS}^{nt}_{n,\mathbb{Y}} = \sum_{k=1}^{K} P_n(X_i = (x^k, x^k) | \Delta X_i = 0) \times \left\| \sqrt{n} \{ F_{n,Y_{i1}|X_i}(y|(x^k, x^k)) - F_{n,Y_{i2}|X_i}(y|(x^k, x^k)) \} \right\|_{\infty,\mathbb{Y}}, \tag{23}$$

$$\widetilde{KS}^{pt}_{n,\mathbb{Y}} = \sum_{k=1}^{K} P_n(X_i = (x^k, x^k) | \Delta X_i = 0) \times \left\| \sqrt{n} \{ F_{n,Y_{i1}|X_i}(y|(x^k, x^k)) - F_{n,Y_{i2}-\lambda_n(x^k)|X_i}(y|(x^k, x^k)) \} \right\|_{\infty,\mathbb{Y}}, \tag{24}$$

$$\widetilde{KS}^{gpt}_{n,\mathbb{Y}} = \sum_{k=1}^{K} P_n(X_i = (x^k, x^k) | \Delta X_i = 0) \times \left\| \sqrt{n} \{ F_{n,Y_{i1}|X_i}(y|(x^k, x^k)) - F_{n,Y_{i2}-\lambda_n(x^k)|X_i}(y|(x^k, x^k)) \} \right\|_{\infty,\mathbb{Y}}, \tag{25}$$

$$\widetilde{CM}^{nt}_{n,\phi} = \sum_{k=1}^{K} P_n(X_i = (x^k, x^k)|\Delta X_i = 0) \times \left\| \sqrt{n}\{F_{n,Y_{i1}|X_i}(y|(x^k, x^k)) - F_{n,Y_{i2}|X_i}(y|(x^k, x^k))\} \right\|_{2,\phi} \tag{26}$$

$$\widetilde{CM}^{pt}_{n,\phi} = \sum_{k=1}^{K} P_n(X_i = (x^k, x^k)|\Delta X_i = 0) \times \left\| \sqrt{n}\{F_{n,Y_{i1}|X_i}(y|(x^k, x^k)) - F_{n,Y_{i2}-\lambda_n|X_i}(y|(x^k, x^k))\} \right\|_{2,\phi} \tag{27}$$

$$\widetilde{CM}^{gpt}_{n,\phi} = \sum_{k=1}^{K} P_n(X_i = (x^k, x^k)|\Delta X_i = 0) \times \left\| \sqrt{n}\{F_{n,Y_{i1}|X_i}(y|(x^k, x^k)) - F_{n,Y_{i2}-\lambda_n(x^k)|X_i}(y|(x^k, x^k))\} \right\|_{2,\phi}, \tag{28}$$

which we will refer to as the disaggregated statistics. In this section, we discuss the simulation results on the relative performance of the aggregated and disaggregated statistics. We consider the same baseline design as in Section 3.3.1. In addition to considering $n = 500, 2000$ as in the baseline design, we let $K = 2, 16$, since the support of X_{it} may have an effect on the relative finite-sample performance of the two categories of statistics. Tables 7-10 report the rejection probabilities of the bootstrap-adjusted statistics using 200 bootstrap replications. Here, we use the standard normal density to compute the CM statistic.

The simulation results for $n = 500$ and $K = 2$ reported in Table 7 indicate that there may be gains in finite-sample power from using the disaggregated version of the KS statistic in lieu of its aggregated version, e.g. see the performance of the *nt* statistic under Model B with $\lambda_2 = 0.25$ as well as the *pt* statistic under Model D with $\sigma_2 = 1.1$. As for the CM statistic, the choice between its aggregated and disaggregated versions seems inconsequential. In terms of size control, there does not seem to be a significant difference between using the aggregated and disaggregated versions. However, this is no longer the case when we increase the support of X_{it} to $K = 16$ while keeping $n = 500$, as illustrated in Table 8. For both the KS and CM statistics, the disaggregated statistics exhibit very poor size control, e.g. see the *nt* statistics under Model A and the *pt* statistics under Model B. It still holds true that the disaggregated KS statistics improve over the aggregated version in terms of finite-sample

power. This may even occur for the CM statistic, e.g. see the nt statistic under Model C. However, the poor size control makes the disaggregated statistics not suitable for this setup.

The simulation results for $n = 2000$ and $K = 2, 16$ are reported in Tables 9 and 10. For $K = 2$ (Table 9), the disaggregated statistics control size quite well. However, the relative gain in finite-sample power from using the disaggregated versions of the KS statistic instead of the aggregated ones is much smaller than when $n = 500$, e.g. see the nt statistic under Model B ($\lambda_2 = 0.25$) and the pt statistic under Model C ($\lambda_2 = 0.25$). For $K = 16$ (Table 10), the results are qualitatively similar to the case when $n = 500$. Even when $n = 2000$, the disaggregated statistics tend to have very poor size control for $K = 16$.

Our simulation results indicate that when n and K are relatively small (e.g. $n = 500$, $K = 2$), the disaggregated statistics for testing time homogeneity may improve finite-sample power over the aggregated statistics while controlling size. However, for large enough K, the disaggregated statistics may suffer from severe size distortions, even when n is relatively large. Since testing time homogeneity will often take place in the context of observational panel data, where the support of the regressor is relatively rich, the disaggregated statistics seem to be relatively unsuitable.

3.3.4 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

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

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 14.[15]

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 $a_t = 0.01, 0.025, 0.05$, which represent 0.02, 0.05 and 0.1 in proportion to the standard deviation of Y_{i1}, respectively. 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 standard deviation of Y_{i1}. Table 11 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 14, 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.

[15] 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 $\xi(x,a,u) = a + (2+a)x + u$.

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 statistics introduced in Section 3.1.2, which we outline in the following

$$KS_{n,\mathbb{Y}}^{pt,T} = \frac{1}{T-1}\sum_{t=1}^{T-1}\left\|\sqrt{n}\left\{F_{n,Y_{it}|\Delta X_{it}}(.|0) - F_{n,Y_{i,t+1}-\Delta\lambda_{n,t}|\Delta X_{it}}(.|0)\right\}\right\|_{\infty,\mathbb{Y}}, \tag{29}$$

$$KS_{n,\mathbb{Y}}^{gpt,T} = \frac{1}{T-1}\sum_{t=1}^{T-1}\left\|\sqrt{n}\left\{F_{n,Y_{it}|\Delta X_{it}}(.|0) - F_{n,Y_{i,t+1}|\Delta X_{it}}(.,\Delta\Lambda_{n,t}|0)\right\}\right\|_{\infty,\mathbb{Y}}, \tag{30}$$

$$CM_{n,\phi}^{pt,T} = \frac{1}{T-1}\sum_{t=1}^{T-1}\left\|\sqrt{n}\left\{F_{n,Y_{it}|\Delta X_{it}}(.|0) - F_{n,Y_{i,t+1}-\Delta\lambda_{n,t}|\Delta X_{it}}(.|0)\right\}\right\|_{2,\phi}, \tag{31}$$

$$CM_{n,\phi}^{gpt,T} = \frac{1}{T-1}\sum_{t=1}^{T-1}\left\|\sqrt{n}\left\{F_{n,Y_{it}|\Delta X_{it}}(.|0) - F_{n,Y_{i,t+1}|\Delta X_{it}}(.,\Delta\Lambda_{n,t}|0)\right\}\right\|_{2,\phi}, \tag{32}$$

where $\Delta\lambda_{n,t} = \sum_{i=1}^{n}(Y_{i,t+1}-Y_{it})1\{\Delta X_{it}=0\}/\sum_{i=1}^{n}1\{\Delta X_{it}=0\}$, $\Delta\lambda_{n,t}(x) = \sum_{i=1}^{n}(Y_{i,t+1}-Y_{it})1\{(X_{it},X_{i,t+1})=(x,x)\}/\sum_{i=1}^{n}1\{(X_{it},X_{i,t+1})=(x,x)\}$, $\Delta\Lambda_{n,t} = \left(\Delta\lambda_{n,t}(x^1),\ldots,\Delta\lambda_{n,t}(x^K)\right)'$. 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 12 reports the rejection probabilities for the bootstrap adjustments of the statistics given above. 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 undersized under Models A-C and inconsistent under Model D in our simulation study. This is in line with what our findings and discussion of the simulation results for $K = 16$ in Section

3.3.3.

Since the choice of the center of the density for the CM statistic is quite important, we also consider an alternative method for demeaning the data that allows us to use densities with zero-mean. We consider the following statistics, where we demean the distribution of Y_{it} for the stayer subsamples as follows

$$CM^{pt,T}_{n,\phi(0)} = \frac{1}{T-1}\sum_{t=1}^{T-1}\left\|\sqrt{n}\left\{F_{n,Y_{it}-\lambda^t_{n,t}|\Delta X_{it}}(.|0) - F_{n,Y_{i,t+1}-\lambda^t_{n,t+1}|\Delta X_{it}}(.|0)\right\}\right\|_{2,\phi}, \quad (33)$$

$$CM^{gpt,T}_{n,\phi(0)} = \frac{1}{T-1}\sum_{t=1}^{T-1}\left\|\sqrt{n}\left\{F_{n,Y_{it}|\Delta X_{it}}(.,\Lambda^t_{n,t}|0) - F_{n,Y_{i,t+1}|\Delta X_{it}}(.,\Lambda^t_{n,t+1}|0)\right\}\right\|_{2,\phi}, \quad (34)$$

where for $\tau = t, t+1$, $\lambda^t_{n,\tau} = \sum_{i=1}^n Y_{i\tau}1\{\Delta X_{it} = 0\}/\sum_{i=1}^n 1\{\Delta X_{it} = 0\}$, $\lambda^t_{n,\tau}(x) = \sum_{i=1}^n Y_{i\tau}1\{(X_{it}, X_{i,t+1}) = (x,x)\}/\sum_{i=1}^n 1\{(X_{it}, X_{i,t+1}) = (x,x)\}$, and $\Lambda^t_{n,\tau} \equiv (\lambda^t_{n,\tau}(x^1), \ldots, \lambda^t_{n,\tau}(x^K))'$.[16] For the above statistics, both empirical cdfs inside the norms in (33) and (34) are centered around zero. Hence, we can use zero-mean densities to compute the CM statistic.

Table 13 reports rejection probabilities of the CM statistics in the above equations using densities $N(0, 0.25)$, $N(0, 0.5)$ and $U(-7, 7)$. For the *pt* statistics, the change in the centering does not seem to affect their performance. Note that (33) both empirical cdfs have zero mean, hence the justification to use zero-mean densities. In (31), the $t+1$ empirical cdf is demeaned to have the same center as the t empirical cdf, and we use 6.5 as the center for the normal densities considered. The reason why using either method does not yield a substantial difference in our study may be attributed to two factors in our design resembling the NLSY sample (see Tables 11 and 14): (1) the mean of the outcome variable does not change substantially from period to period, (2) the proportion of stayer subpopulations is quite high.

The *gpt* statistics however seem to improve substantially in terms of their consistency under Model D with $|\sigma_{t+1} - \sigma_t| = 0.05, 0.1$, which are 5% and 10% of standard deviation.

[16] Since we compare several two-period combinations (t and $t+1$, for $t = 1, \ldots, T-1$), the stayer subpopulations ($\Delta X_{it} = 0$) change depending on the two periods we are comparing. Hence, the superscript t that we add to the time trends reflect this issue.

This finding may be explained by the following. Note that under the null hypothesis of time homogeneity with generalized time trends, stayer subpopulations, i.e. $(X_{it}, X_{i,t+1}) = (x, x)$, with different x have different time trends, $\lambda_t(x)$ and $\lambda_{t+1}(x)$. Hence, the mean of the distribution of each of these subpopulations in the base period t, i.e. $E[Y_{it}|(X_{it}, X_{i,t+1}) = (x, x)]$, may be substantially different from the overall mean of all stayers, $E[Y_{it}|\Delta X_{it} = 0]$. Demeaning as in (32) demeans the $t+1$ distribution of each subpopulation $(X_{it}, X_{i,t+1}) = (x, x)$ so that it has the same mean of the base period $E[Y_{it}|(X_{it}, X_{i,t+1}) = (x, x)]$, whereas the demeaning in (34) ensures that the distributions at t and $t+1$ are both centered around zero. Hence, all stayer subpopulations are centered around the same mean before we aggregate over them, and we use this same mean as the center of the density for the CM statistic. This improvement in centering is naturally expected to improve the statistic's ability to detect time heterogeneity in the scale.

4 Empirical Illustration: Returns to Schooling

4.1 Revisiting Angrist and Newey (1991)

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

$$Y_{it} = X_{it}'\beta + \mathcal{A}_i + \mathcal{U}_{it}, \tag{35}$$

$$E[\mathcal{U}_{it}|X_{i1}, \ldots, X_{it}, \mathcal{A}_i] = 0 \tag{36}$$

where X_{it} and β are $p \times 1$ vectors.[17] The key identifying assumption in the above is the restriction in (36), which is referred to by Chamberlain (1984) as "strict exogeneity conditional on a latent variable." Chamberlain (1984) proposes a minimun chi-square (MCS) procedure to test the restrictions implied by (35) and (36). Angrist and Newey (1991) show that the over-identification tests from three stage least squares (3SLS) equivalents of the Chamber-

[17] To clarify why we choose to define p, and not K, as the dimension of X_{it}, consider two settings, one where a scalar X_{it} ($p = 1$) can take four possible values and another where $p = 2$ and both elements of X_{it} are binary. In both situations, $K = 4$, even though p is different.

lain (1984) procedure yield statistics that are identical to the MCS statistic.[18] The testable restrictions are obtained from considering the linear prediction of $\mathcal{A}_i$ given X_i

$$\mathcal{A}_i = X_{i1}'\eta_1 + \cdots + X_{iT}'\eta_T + \mathcal{E}_i. \tag{37}$$

By construction, $\mathcal{E}_i$ is uncorrelated with $X_{i1}, \ldots, X_{iT}$.[19] Taking deviations of Y_{it} from Y_{iT} for $t = 1, \ldots, T-1$ and plugging (37) 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.[20] Their over-identification test for the union wage effects 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. 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 14 and are quite similar to their counterparts reported in Angrist and Newey (1991, Table1). The main ANACOVA specification of Mincer's human capital earnings function in Angrist and Newey (1991) is

[18] 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.

[19] Chamberlain (1984) points out that (37) 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.

[20] It is important to note here that 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.

given in the following

$$Y_{it} = \lambda_t + \beta_1 S_{it} + \beta_2 S_{it}^2 + \beta_3 Age_{it}^2 + \beta_4 S_{it} * Age_{it}^2 + \beta_5 Union_{it} + \mathcal{A}_i + \mathcal{U}_{it} \tag{38}$$

where Y is log earnings and S is years of completed education (highest grade completed). We replicate the ANACOVA estimates of the above equation in Table 15 for the full sample as well as for the period 1983-1984. 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 $Grade$, Age^2 and $Union$.[21] All of the specifications considered in Angrist and Newey (1991) are nonlinear in schooling. However, they are not nonlinear in the relationship between schooling and unobservables, $\mathcal{A}_i$ and $\mathcal{U}_{it}$. In the following section, we will revisit this application and test for time homogeneity in order to examine whether the APE of schooling is nonparamterically 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 the following model exhibiting time homogeneity up to a parallel trend, which imposes no functional form restrictions on the relationship between schooling and unobservables in the human capital earnings function,

$$Y_{it} = \xi(S_{it}, \mathcal{A}_i, \mathcal{U}_{it}) + \lambda_t,$$
$$\mathcal{U}_{it}|S_i, \mathcal{A}_i \overset{d}{=} \mathcal{U}_{i1}|S_i, \mathcal{A}_i.$$

[21] 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 $Grade$ in RF, which is similar to the estimate in Angrist and Newey (1991) when we use 1983-1984 years, but not when we use the full revised sample. For details on the reasoning behind the different specification, see Angrist and Newey (1991).

where Y_{it} is log earnings and S_{it} is schooling measured by the highest grade completed, previously referred to by *Grade*.

The model implies that individuals that do not change their schooling status, i.e. stayers, should have the same distribution for log earnings across time, once appropriately demeaned. Formally, the testable implication is given as follows

$$H_0 : F_{Y_{it}|\Delta S_{it}}(.|0) = F_{Y_{it}-\Delta\lambda_{t+1}^t|\Delta S_{it}}(.|0), \text{ for } t = 1, \ldots, T-1 \tag{39}$$

To test the above null, Table 16 reports the p-values for the bootstrap-adjusted KS and CM statistics for the full sample (1983-1987) using the statistics in (29) and (31), where we substitute $X_i = S_i$. 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% level of significance.

In addition, Table 16 reports the p-values for the bootstrap-adjusted statistics for two-period combinations, 1983-84, 1984-85, 1985-86, 1986-87, for the KS and CM statistics given in (14) and (18), where $X_i = (S_{it}, S_{i,t+1})$. The p-value of the F test reported in the same table for every two-period combination we consider is based on another implication of time homogeneity with a parallel trend, which is given by

$$E[Y_{i,t+1} - Y_{it}|(S_{it}, S_{i,t+1}) = (s, s)] = E[Y_{i,t+1} - Y_{it}|(S_{it}, S_{i,t+1}) = (s', s')] \ \ \forall s, s' \in \mathbb{S}, s \neq s', \tag{40}$$

where $\mathbb{S}$ denotes the support of S_{it}. Most of the p-values of the statistics for 1983-84 and 1984-85 are smaller than the statistics for the full sample, whereas for 1985-86 and 1986-87 the p-values are larger than the ones for the full sample. Using the KS and $CM(6.5, 0.5)$for 1984-85, time homogeneity is not rejected only marginally at the 10% level. The F-statistic for this period combination rejects the implication in (40) at the 10% level of significance.

Table 19 reports the APE for movers for every two-period combination considered above.

The APEs for movers $(S_{it}, S_{i,t+1}) = (s, s+1)$ is estimated as follows

$$\begin{aligned}&\hat{\beta}_t(s \to s+1|(S_{it}, S_{i,t+1}) = (s, s+1)) \\ &= \frac{\sum_{i=1}^n (Y_{i,t+1} - Y_{it}) 1\{(S_{it}, S_{i,t+1}) = (s, s+1)\}}{\sum_{i=1}^n 1\{(S_{it}, S_{i,t+1}) = (s, s+1)\}} - \Delta\lambda_{n,t},\end{aligned} \tag{41}$$

where $\Delta\lambda_{n,t} = \sum_{i=1}^n (Y_{i,t+1} - Y_{it}) 1\{\Delta S_{it} = 0\} / \sum_{i=1}^n 1\{\Delta S_{it} = 0\}$. The APE for all movers[22] is given by

$$\begin{aligned}&\hat{\beta}_t(\Delta S_i = 1) \\ &= \sum_{s \in \mathbb{S}} P_n((S_{it}, S_{i,t+1}) = (s, s+1)|\Delta S_{it} = 1) \hat{\beta}(s \to s+1|(S_{it}, S_{i,t+1}) = (s, s+1)).\end{aligned} \tag{42}$$

The standard errors are computed under the assumption of cross-sectional independence. Our APE estimates are not significant, except in 1985-1986. Our results provide evidence against the implication of the linear model that responses to changes in schooling status are homogeneous. We find evidence of heterogeneity in responses to schooling which may be attributed to two possible channels. The first channel would be that changes in schooling are heterogeneous in the starting level of schooling. This is in line with the empirical evidence quoted in Card (1999) that changes in schooling status have nonlinear effects on earnings that are particularly significant around the completion of terminal degrees. We find evidence thereof in our results. For instance, the significance of the APE in 1985-86 is driven solely by individuals who completed their bachelor degree, i.e. 16 years of schooling. Another channel of response heterogeneity is due to the nonseparability between schooling and unobservables, $\mathcal{A}_i$ and $\mathcal{U}_{it}$. Since movers vary from year-to-year and their responses are heterogeneous, our model would not predict that the APE of schooling is constant over time, even under time homogeneity.

Our estimates of the APE indicate that the linear model may be misspecified. Combining this finding with our inability to reject time homogeneity, we conjecture that the rejection

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

of the over-identification test proposed by Angrist and Newey (1991) is largely due to the misspecification of the linear model. We conclude that the APE of schooling on earnings may be identified nonparametrically by time homogeneity in the NLSY sample 1983-1987.

When we test nonparametric identification, we test whether the data can identify the effect of interest. When we use over-identification tests such as Angrist and Newey (1991), we test whether a parametric model imposed on the data can identify the effect of interest. The well-known shortcoming of the latter approach is that when the model is misspecified, we reject the test, even if the data can identify the effect of interest, i.e. the effect is nonparametrically identified. This empirical application illustrates this issue and points to the merits of testing nonparametric identification.

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

$$\begin{aligned} E[Y_{i2} - Y_{i1}|X_i = (x,x')] &= E[Y_{i2} - Y^x_{i2}|X_i = (x,x')] + E[Y^x_{i2} - Y_{i1}|X_i = (x,x')] \\ &= \underbrace{\beta_2(x \to x'|X_i = (x,x'))}_{\text{APE for Period 2 for } X_i = (x,x')} + \underbrace{E[Y^x_{i2} - Y_{i1}|X_i = (x,x')]}_{\text{Counterfactual Trend}}. \end{aligned} \tag{43}$$

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

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

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) Without loss of generality, we can normalize $\lambda_1(x) = 0$ for all $x \in \mathbb{X}$. We can thus drop the subscript from $\lambda_2(x)$ in the following. Plugging in for $\xi_2(x,a,u) = \xi(x,a,u) + \lambda(x)$, the condition in Theorem 2.1 simplifies to the following

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

Clearly, $\lambda(x)$ cancels out from the above integrals and the above simplifies to

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

Now we show that $\mathcal{U}_{i1}|X_i, \mathcal{A}_i \overset{d}{=} \mathcal{U}_{i2}|X_i, \mathcal{A}_i$ implies that $\forall \underline{x} \in \{(x,x),(x,x')\}$,[23]

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

By Assumption 2.1(iii), Fubini's theorem applies and the above yields

$$\int \left(\int (\xi(x,a,u_2) - \xi(x,a,u_1)) f_{\mathcal{U}_{i1},\mathcal{U}_{i2}|X_i,\mathcal{A}_i}(u_1,u_2|\underline{x},a) d(u_1,u_2) \right) f_{\mathcal{A}_i|X_i}(a|\underline{x}) da = 0. \tag{44}$$

Note that by definition,

$$f_{\mathcal{U}_{i1},\mathcal{U}_{i2}|X_i,\mathcal{A}_i} = f_{\mathcal{U}_{i1}|X_i,\mathcal{A}_i,\mathcal{U}_{i2}} f_{\mathcal{U}_{i2}|X_i,\mathcal{A}_i} = f_{\mathcal{U}_{i2}|X_i,\mathcal{A}_i,\mathcal{U}_{i1}} f_{\mathcal{U}_{i1}|X_i,\mathcal{A}_i} = f_{\mathcal{U}_{i2}|X_i,\mathcal{A}_i,\mathcal{U}_{i1}} f_{\mathcal{U}_{i2}|X_i,\mathcal{A}_i}$$

where the last equality follows from $\mathcal{U}_{i1}|X_i, \mathcal{A}_i \overset{d}{=} \mathcal{U}_{i2}|X_i, \mathcal{A}_i$. By Assumption 2.2, $f_{\mathcal{U}_{i2}|X_i,\mathcal{A}_i} > 0$, so

[23] 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 $f_{\mathcal{A}_i,\mathcal{U}_{i1},\mathcal{U}_{i2}|X_i}(.,.,.|(x,x')) \neq f_{\mathcal{A}_i,\mathcal{U}_{i1},\mathcal{U}_{i2}|X_i}(.,.,.|(x,x))$.

we can divide the above equation by it, which yields $f_{\mathcal{U}_{i1}|X_i,\mathcal{A}_i,\mathcal{U}_{i2}} = f_{\mathcal{U}_{i2}|X_i,\mathcal{A}_i,\mathcal{U}_{i1}}$. It follows that

$$\int \xi(x,a,u_1) f_{\mathcal{U}_{i1}|X_i,\mathcal{A}_i,\mathcal{U}_{i2}}(u_1|\underline{x},a,u_2)du_1 = \int \xi(x,a,u_2) f_{\mathcal{U}_{i2}|X_i,\mathcal{A}_i,\mathcal{U}_{i1}}(u_2|\underline{x},a,u_1)du_2$$

Now integrating the above with respect to $f_{\mathcal{U}_{i1}|X_i,\mathcal{A}_i} = f_{\mathcal{U}_{i2}|X_i,\mathcal{A}_i}$, implies that

$$\begin{aligned}
&\int \left(\int \xi(x,a,u_1) f_{\mathcal{U}_{i1}|X_i,\mathcal{A}_i,\mathcal{U}_{i2}}(u_1|\underline{x},a,u_2)du_1 \right) f_{\mathcal{U}_{i1}|X_i,\mathcal{A}_i}(u_1|\underline{x},a)du_1 \\
&= \int \left(\int \xi(x,a,u_2) f_{\mathcal{U}_{i2}|X_i,\mathcal{A}_i,\mathcal{U}_{i1}}(u_2|\underline{x},a,u_1)du_2 \right) f_{\mathcal{U}_{i2}|X_i,\mathcal{A}_i}(u_2|\underline{x},a)du_2.
\end{aligned}$$

By Fubini's theorem, the above implies that the integrand in (44) is zero and the result follows.
(ii) is straightforward from the following:

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

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 $Y_{it}^x - \lambda_t(x) = \xi(x,a,u)$ 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

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

□

A.2 Proofs of Section 3 Results

Proof (Theorem 3.1)
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.

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

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

Now note that

$$F_{2,n}(.|\Delta X_i = 0) = \sum_{k=1}^{K} P_n(X_{i1} = X_{i2} = x^k) F_{2,n}(. + \lambda(x)|X_{i1} = X_{i2} = x^k)$$
$$F_2(.|\Delta X_i = 0) = \sum_{k=1}^{K} P(X_{i1} = X_{i2} = x^k) F_2(. + \lambda(x)|X_{i1} = X_{i2} = x^k) \tag{45}$$

Since Assumption 3.1 holds, it follows that, for $x \in \mathbb{X}$, $F_2(. + \lambda(x)|X_{i1} = X_{i2} = x)$ is Hadamard differentiable tangentially to $\mathbb{D} \times \mathbb{Y}$ by Lemma B.1 in Section B.3 the supplementary appendix, where $\mathbb{D} = \{g \in \mathcal{L}^\infty(\mathcal{F}) : \text{g is } \rho_2\text{-uniformly continuous}\}$, where ρ_2 is the variance metric. The Hadamard derivative is given by $\phi'_{F_2(.|x),\lambda(x)}(g, \epsilon) = g(.+\lambda(x)) + \epsilon f_2(.+\lambda(x)|X_{i1} = X_{i2} = x)$, where the subscript $F_2(.|x)$ denotes $F_2(.|X_{i1} = X_{i2} = x)$. Now $F_1(.|X_{i1} = X_{i2} = x) - F_2(. + \lambda(x)|X_{i1} = X_{i2} = x)$ is trivially Hadamard differentiable tangentially to $\mathbb{D} \times \mathcal{L}^\infty(\mathbb{Y}) \times \mathbb{Y}$.

Let $F_{t|x}(.) = F_t(.|X_{i1} = X_{i2} = x)$ and $F_{t|x,n}(.)$ be the sample analogue thereof. Now noting that

$$\sqrt{n} \begin{pmatrix} F_{1|.,n}(.) - F_{1|.}(.) \\ F_{2|.,n}(.) - F_{2|.}(.) \\ \Lambda_n - \Lambda \end{pmatrix} \rightsquigarrow \begin{pmatrix} \mathbb{G}_{1|.} \\ \mathbb{G}_{2|.} \\ \mathcal{E} \end{pmatrix} \tag{46}$$

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. For the following, we define $\mathcal{E}(x)$ as follows, $\sqrt{n}(\lambda_n(x) - \lambda(x)) \xrightarrow{p} \mathcal{E}(x)$.

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

$$\sqrt{n}\left(F_{1|.,n}(.) - F_{2|.,n}(. + \lambda_n(.)) - (F_{1|.}(.) - F_{2|.}(. + \lambda(.)))\right) \mapsto \mathbb{G}_{1,2|.}, \tag{47}$$

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

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

$$\sqrt{n}\left(\hat{F}_{1|.,n}(.) - \hat{F}_{2|.,n}(. + \hat{\lambda}_n(.)) - (F_{1|.,n}(.) - F_{2|.,n}(. + \lambda_n(.)))\right) \mapsto \mathbb{G}_{1,2|.}. \tag{48}$$

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 (46) 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 $\sup_{f \in \mathcal{F}} |\int (f - \int f dF_t(y|X_{i1} = X_{i2} = x))^2 dF_t(y|X_{i1} = X_{i2} = x)| < \infty$, the conditions in Theorem 3.6.2. hold. Thus, (c) holds, which implies (48).

Now we relate (47) and (48) to $\mathbb{G}_{n|\Delta X_i=0}$ and $\hat{\mathbb{G}}_{n|\Delta X_i=0}$, respectively. Let $F_{1|x,n}(.) \equiv F_{1,n}(.|X_{i1} =$

$X_{i2} = x$).

$$
\begin{aligned}
&\mathbb{G}_{n|\Delta X_i=0} \\
&= \sqrt{n}(F_{1,n}(.|\Delta X_i) - F_{2,n}(., \Lambda_n|\Delta X_i = 0)) - \sqrt{n}(F_1(.|\Delta X_i = 0) - F_2(., \Lambda|\Delta X_i = 0)) \\
&= \sqrt{n}\sum_{k=1}^{K} P_n(X_{i1} = X_{i2} = x^k|\Delta X_i = 0)\left(F_{1|x^k,n}(.) - F_{2|x^k,n}(. + \lambda_n(x^k))\right) \\
&- \sqrt{n}\sum_{k=1}^{K} P_n(X_{i1} = X_{i2} = x^k|\Delta X_i = 0)\left(F_{1|x^k}(.) - F_{2|x^k}(. + \lambda(x^k))\right) \\
&= \sqrt{n}\sum_{k=1}^{K} P(X_{i1} = X_{i2} = x^k|\Delta X_i = 0)\left(F_{1|x^k,n}(.) - F_{2|x^k,n}(. + \lambda_n(x^k))\right) \\
&- \sqrt{n}\sum_{k=1}^{K} P(X_{i1} = X_{i2} = x^k|\Delta X_i = 0)\left(F_{1|x^k}(.) - F_{2|x^k}(. + \lambda(x^k))\right) \\
&+ \sum_{k=1}^{K}(P_n(X_{i1} = X_{i2} = x^k|\Delta X_i = 0) - P(X_{i1} = X_{i2} = x^k|\Delta X_i = 0)) \\
&\times \sqrt{n}\left(F_{1|x^k,n}(.) - F_{2|x^k,n}(. + \lambda_n(x^k)) - (F_{1|x^k}(.) - F_{2|x^k}(. + \lambda(x^k)))\right) \qquad (49) \\
&\rightsquigarrow \sum_{k=1}^{K} P(X_{i1} = X_{i2} = x^k|\Delta X_i = 0)(\mathbb{G}_{1|x^k} - \phi'_{F_2,\lambda(x^k)}(\mathbb{G}_{2|x^k}, \mathcal{E}(x^k))) \\
&\equiv \mathcal{H}(\Lambda) \qquad (50)
\end{aligned}
$$

since the last term of the last equality converges to zero in probability by the weak convergence of (49) and $P_n(X_{i1} = X_{i2} = x^k|\Delta X_i = 0) \xrightarrow{p} P(X_{i1} = X_{i2} = x^k|\Delta X_i = 0)$ for all $k = 1, 2, ..., K$. Hence, we have shown that the empirical process, $\mathbb{G}_{n|\Delta X_i=0}$ converges to a tight Brownian bridge.

Now the bootstrap empirical process can be decomposed as follows

$$
\begin{aligned}
&\hat{\mathbb{G}}_{n|\Delta X_i=0} \\
&= \sqrt{n}\left(\hat{F}_{1,n}(.|\Delta X_i = 0) - \hat{F}_{2,n}(., \hat{\Lambda}_n|\Delta X_i = 0) - \{F_{1,n}(.|\Delta X_i = 0) - F_{2,n}(., \Lambda_n|\Delta X_i = 0)\}\right) \\
&= \sqrt{n}\sum_{k=1}^{K} \hat{P}_n(X_{i1} = X_{i2} = x^k|\Delta X_i = 0) \\
&\times \left(\hat{F}_{1|x^k,n}(. + \hat{\lambda}_n(x^k)) - \hat{F}_{2|x^k,n}(.) - (F_{1|x^k,n}(. + \lambda_n(x^k)) - F_{2|x^k,n}(.))\right) \\
&= \sqrt{n}\sum_{k=1}^{K} P(X_{i1} = X_{i2} = x^k|\Delta X_i = 0) \\
&\times \left(\hat{F}_{1|x^k,n}(.) - \hat{F}_{2|x^k,n}(. + \hat{\lambda}_n(x^k)) - (F_{1|x^k,n}(.) - F_{2|x^k,n}(. + \lambda_n(x^k)))\right) \\
&+ \sum_{k=1}^{K}(\hat{P}_n(X_{i1} = X_{i2} = x^k|\Delta X_i = 0) - P(X_{i1} = X_{i2} = x^k|\Delta X_i = 0)) \\
&\times \sqrt{n}\left(\hat{F}_{1|x^k,n}(.) - \hat{F}_{2|x^k,n}(. + \hat{\lambda}_n(x^k)) - (F_{1|x^k,n}(.) - F_{2|x^k,n}(. + \lambda_n(x^k)))\right) \\
&\rightsquigarrow \mathcal{H}(\Lambda) \qquad (51)
\end{aligned}
$$

where the first term of the last equality follows by the continuous mapping theorem and (48). The second term converges to zero by (48) and $(\hat{P}_n(X_{i1} = X_{i2} = x^k|\Delta X_i = 0) - P(X_{i1} = X_{i2} = x^k|\Delta X_i = 0)) \xrightarrow{p} 0$ by Lemma B.2 in Section B.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 $F_1(.|\Delta X_i = 0)$ and $F_2(.|\Delta X_i = 0)$ 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

$$\hat{c}_n^{KS} = \inf\{t : \hat{P}_n(\|\hat{\mathbb{G}}_{n|\Delta X_i=0}\|_{\infty,\mathbb{Y}} > t) \leq \alpha\},$$
$$\hat{c}_n^{CM} = \inf\{t : \hat{P}_n(\|\hat{\mathbb{G}}_{n|\Delta X_i=0}\|_{2,\phi} > t) \leq \alpha\},$$

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

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

Thus, under the null, the bootstrap-adjusted statistics have 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 $KS_{n,\mathbb{Y}}(\Lambda_n)$ and $CM_{n,\phi}(\Lambda_n)$ clearly diverge to infinity under an alternative hypothesis. Thus, the tests are consistent against any fixed alternative, which proves (ii). □

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 $m_l = |\{\underline{x} \in \mathbb{X}_l : \underline{x} \in \mathbb{X}\}|$, where $|S|$ denotes the cardinality of a set S. Our statistics can be written as follows:

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

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

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

$$\begin{pmatrix} \sqrt{n}\left(F_{1,n}(.|X_i = \underline{x}) - F_1(.|X_i = \underline{x})\right)_{\underline{x}\in\mathbb{X}^T} \\ \sqrt{n}\left(F_{2,n}(.|X_i = \underline{x}) - F_2(.|X_i = \underline{x})\right)_{\underline{x}\in\mathbb{X}^T} \\ \cdots \\ \sqrt{n}\left(F_{T,n}(.|X_i = \underline{x}) - F_T(.|X_i = \underline{x})\right)_{\underline{x}\in\mathbb{X}^T} \end{pmatrix} \rightsquigarrow \mathbb{G} \tag{52}$$

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.

$$\left(\begin{pmatrix} \sqrt{n}(F_{1,n}(.|\underline{x}_j) - \bar{F}_{1,n,l}(.) - (F_1(.|\underline{x}_j) - \bar{F}_{1,l}(.))) \\ \cdots \\ \sqrt{n}(F_{T,n}(.|\underline{x}_j) - \bar{F}_{T,n,l}(.) - (F_T(.|\underline{x}_j) - \bar{F}_{T,l}(.))) \end{pmatrix}_{j=1,2,\ldots,m_l}\right)_{l=1,2,\ldots,L} \rightsquigarrow \mathcal{H},$$

where $\mathcal{H} \equiv ((\mathcal{H}_{j,l})_{j=1,..,m_l})_{l=1,..,L}$. Note that the above process is a $(T\sum_{l=1}^{L} m_l)\times 1$ 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).

$$\left(\left(\begin{array}{c} \sqrt{n}(\hat{F}_{1,n}(.|\underline{x}_j) - \hat{\bar{F}}_{1,n,l}(.) - (F_{1,n}(.|\underline{x}_j) - \bar{F}_{1,n,l}(.))) \\ \cdots \\ \sqrt{n}(\hat{F}_{T,n}(.|\underline{x}_j) - \hat{\bar{F}}_{T,n,l}(.) - (F_{T,n}(.|\underline{x}_j) - \bar{F}_{T,n,l}(.))) \end{array} \right)_{j=1,2,\ldots,m_l} \right)_{l=1,2,\ldots,L}$$
$$\rightsquigarrow \mathcal{H}.$$

Now we give the limiting statistics as follows,

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

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

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 the $P(KS_{\mathbb{Y}} = 0) = 0$ 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

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

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 $(P_n(X_i = \underline{x}|h(X_i) = h_l) - P(X_i = \underline{x}|h(X_i) = h_l)) \xrightarrow{p} 0$ 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. □

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 Richard Carson, Sarojini Hirshleifer, Philip Neary, Leah Nelson, Kevin Song, seminar participants at UC Berkeley, UCSD, UCSB, UC Davis, NYU Abu Dhabi, and Georgetown Qatar for helpful comments. Finally, I am especially grateful to Prashant Bharadwaj for sharing the dataset used here.

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., 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): "Robust Priors in Nonlinear Panel Data Models," *Econometrica*, 77(2), 489–536.

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 (2014): "Nonparametric Identification in Panels using Quantiles," *Cemmap Working Paper*.

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.

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*.

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 ($\lambda_2 = 0.25$)												
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 B ($\lambda_2 = 0.5$)												
nt	0.935	0.971	0.990	0.995	0.998	0.999	1.000	1.000	1.000	1.000	1.000	1.000
pt	0.014	0.025	0.054	0.023	0.048	0.096	0.024	0.038	0.071	0.032	0.058	0.109
gpt	0.001	0.010	0.030	0.015	0.031	0.063	0.008	0.028	0.047	0.027	0.056	0.094
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 ($\lambda_2 = 0.25$)												
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 C ($\lambda_2 = 0.5$)												
nt	0.923	0.968	0.985	0.900	0.948	0.975	1.000	1.000	1.000	1.000	1.000	1.000
pt	0.966	0.984	0.992	0.942	0.967	0.981	1.000	1.000	1.000	1.000	1.000	1.000
gpt	0.001	0.010	0.030	0.016	0.029	0.068	0.009	0.022	0.050	0.033	0.053	0.101
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 ($\lambda_2 = 0.5$, $\sigma_2 = 1.1$)												
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
Model D ($\lambda_2 = 0.5$, $\sigma_2 = 1.2$)												
nt	1.000	1.000	1.000	0.944	0.971	0.995	1.000	1.000	1.000	1.000	1.000	1.000
pt	0.975	0.988	0.996	0.397	0.521	0.665	1.000	1.000	1.000	0.995	1.000	1.000
gpt	0.050	0.095	0.173	0.253	0.349	0.456	0.570	0.714	0.833	0.907	0.948	0.982
cre	0.035	0.061	0.114	0.025	0.053	0.103	0.035	0.060	0.108	0.029	0.042	0.085

Notes: The table reports the rejection probabilities across $1,000$ simulations for the bootstrap adjustments for the statistics defined in Equations (13)-(20), 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: Simulation Results on the Choice of Density in the CM Statistic (Demeaned Y_{i2}): $n = 500$, $T = K = 2$

ϕ	N(0,1)			N(0,3)			U(-2,2)			U(-5,5)		
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	0.026	0.050	0.106	0.025	0.053	0.110	0.029	0.052	0.112	0.025	0.053	0.111
pt	0.022	0.046	0.097	0.022	0.050	0.101	0.023	0.049	0.097	0.024	0.053	0.094
gpt	0.017	0.030	0.071	0.017	0.026	0.066	0.014	0.035	0.070	0.018	0.031	0.066
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 ($\lambda_2 = 0.25$)												
nt	0.348	0.464	0.593	0.547	0.644	0.762	0.345	0.450	0.578	0.569	0.662	0.772
pt	0.022	0.044	0.094	0.022	0.047	0.096	0.021	0.047	0.094	0.024	0.048	0.090
gpt	0.017	0.029	0.069	0.014	0.027	0.066	0.019	0.034	0.073	0.019	0.030	0.067
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 ($\lambda_2 = 0.5$)												
nt	0.938	0.964	0.986	0.996	0.999	0.999	0.932	0.965	0.984	0.998	0.999	1.000
pt	0.019	0.047	0.099	0.023	0.050	0.090	0.022	0.050	0.096	0.024	0.048	0.087
gpt	0.015	0.030	0.070	0.016	0.029	0.062	0.016	0.033	0.074	0.023	0.031	0.064
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 ($\lambda_2 = 0.25$)												
nt	0.272	0.369	0.513	0.521	0.611	0.732	0.275	0.354	0.508	0.544	0.642	0.729
pt	0.341	0.450	0.573	0.617	0.698	0.780	0.354	0.451	0.564	0.625	0.703	0.783
gpt	0.014	0.025	0.076	0.012	0.032	0.062	0.018	0.038	0.072	0.014	0.034	0.068
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 ($\lambda_2 = 0.5$)												
nt	0.907	0.949	0.976	0.995	0.997	1.000	0.853	0.916	0.954	0.996	0.998	0.999
pt	0.946	0.967	0.984	0.998	1.000	1.000	0.917	0.950	0.970	0.998	0.999	1.000
gpt	0.018	0.030	0.073	0.013	0.031	0.063	0.019	0.037	0.068	0.016	0.037	0.069
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 ($\lambda_2 = 0.5$, $\sigma_2 = 1.1$)												
nt	0.932	0.958	0.982	1.000	1.000	1.000	0.897	0.948	0.972	1.000	1.000	1.000
pt	0.102	0.160	0.269	0.631	0.703	0.797	0.106	0.163	0.240	0.700	0.777	0.851
gpt	0.068	0.128	0.201	0.078	0.128	0.208	0.094	0.145	0.214	0.080	0.124	0.205
cre	0.026	0.054	0.100	0.030	0.058	0.100	0.029	0.052	0.104	0.037	0.061	0.100
Model D ($\lambda_2 = 0.5$, $\sigma_2 = 1.2$)												
nt	0.947	0.972	0.991	1.000	1.000	1.000	0.916	0.951	0.980	1.000	1.000	1.000
pt	0.393	0.517	0.666	0.997	0.999	1.000	0.322	0.435	0.555	0.999	1.000	1.000
gpt	0.259	0.347	0.467	0.321	0.398	0.526	0.287	0.382	0.505	0.333	0.425	0.544
cre	0.021	0.051	0.098	0.028	0.058	0.108	0.030	0.054	0.103	0.033	0.063	0.104

Notes: The table reports the rejection probabilities across $1,000$ simulations for the bootstrap adjustments for the statistics defined in Equations (17)-(20), where *nt*, *pt*, *gpt*, and *cre* follow the convention of the superscripts in the equations. ϕ denotes the density used to compute the CM statistic. Bold font indicates that the model considered satisfies the null hypothesis for the statistic in question. Models A-D are defined in Table 1.

Table 4: Simulation Results on the Choice of Density in the CM Statistic (Demeaned Y_{i2}): $n = 2000$, $T = K = 2$

ϕ	N(0,1)			N(0,3)			U(-2,2)			U(-5,5)		
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	0.029	0.057	0.108	0.036	0.071	0.113	0.024	0.053	0.107	0.039	0.070	0.112
pt	0.034	0.062	0.109	0.039	0.074	0.118	0.029	0.056	0.105	0.037	0.069	0.108
gpt	0.032	0.050	0.102	0.027	0.043	0.081	0.026	0.048	0.102	0.028	0.041	0.083
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 ($\lambda_2 = 0.25$)												
nt	0.945	0.973	0.994	0.997	0.998	1.000	0.936	0.962	0.988	0.998	0.999	1.000
pt	0.032	0.064	0.114	0.041	0.067	0.115	0.030	0.057	0.109	0.039	0.063	0.115
gpt	0.031	0.056	0.096	0.027	0.048	0.078	0.033	0.051	0.101	0.025	0.045	0.086
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 ($\lambda_2 = 0.5$)												
nt	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
pt	0.032	0.065	0.110	0.037	0.061	0.114	0.029	0.057	0.105	0.040	0.062	0.111
gpt	0.034	0.056	0.100	0.028	0.049	0.084	0.029	0.050	0.100	0.029	0.050	0.088
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 ($\lambda_2 = 0.25$)												
nt	0.927	0.959	0.983	0.992	0.997	0.999	0.910	0.943	0.967	0.994	0.996	0.999
pt	0.967	0.983	0.993	0.998	0.999	1.000	0.946	0.968	0.984	0.998	0.999	1.000
gpt	0.031	0.051	0.103	0.028	0.045	0.082	0.030	0.048	0.098	0.026	0.044	0.089
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 ($\lambda_2 = 0.5$)												
nt	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
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	0.027	0.055	0.093	0.026	0.046	0.079	0.025	0.045	0.100	0.026	0.046	0.080
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 ($\lambda_2 = 0.5$, $\sigma_2 = 1.1$)												
nt	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
pt	0.560	0.690	0.801	0.999	1.000	1.000	0.451	0.566	0.682	1.000	1.000	1.000
gpt	0.317	0.424	0.555	0.395	0.498	0.609	0.375	0.464	0.580	0.420	0.518	0.635
cre	0.027	0.046	0.084	0.031	0.053	0.095	0.022	0.042	0.077	0.038	0.061	0.104
Model D ($\lambda_2 = 0.5$, $\sigma_2 = 1.2$)												
nt	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
pt	0.994	1.000	1.000	1.000	1.000	1.000	0.942	0.962	0.984	1.000	1.000	1.000
gpt	0.898	0.947	0.981	0.973	0.986	0.992	0.946	0.968	0.984	0.975	0.989	0.994
cre	0.030	0.046	0.080	0.031	0.055	0.095	0.027	0.039	0.078	0.034	0.062	0.107

Notes: The table reports the rejection probabilities across $1,000$ simulations for the bootstrap adjustments for the statistics defined in Equations (17)-(20), where *nt*, *pt*, *gpt*, and *cre* follow the convention of the superscripts in the equations. ϕ denotes the density used to compute the CM statistic. Bold font indicates that the model considered satisfies the null hypothesis for the statistic in question. Models A-D are defined in Table 1.

Table 5: Simulation Results on the Choice of Density in the CM Statistic (Demeaned Y_{i1} and Y_{i2}): $n = 500$, $T = K = 2$

ϕ	N(0,1)			N(0,3)			U(-2,2)			U(-5,5)		
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												
pt	0.003	0.012	0.034	0.008	0.016	0.034	0.007	0.019	0.043	0.007	0.018	0.042
gpt	0.015	0.036	0.063	0.015	0.030	0.060	0.015	0.033	0.063	0.016	0.031	0.055
Model B ($\lambda_2 = 0.25$)												
pt	0.003	0.010	0.036	0.007	0.019	0.036	0.006	0.021	0.044	0.010	0.021	0.041
gpt	0.021	0.034	0.061	0.018	0.034	0.060	0.017	0.034	0.064	0.018	0.034	0.060
Model B ($\lambda_2 = 0.5$)												
pt	0.003	0.010	0.037	0.005	0.018	0.039	0.008	0.022	0.044	0.005	0.021	0.043
gpt	0.019	0.033	0.061	0.016	0.030	0.056	0.015	0.033	0.060	0.016	0.030	0.054
Model C ($\lambda_2 = 0.25$)												
pt	0.139	0.234	0.361	0.509	0.595	0.682	0.176	0.252	0.372	0.529	0.619	0.692
gpt	0.019	0.032	0.061	0.017	0.029	0.058	0.016	0.032	0.062	0.017	0.029	0.056
Model C ($\lambda_2 = 0.5$)												
pt	0.757	0.853	0.929	0.996	0.997	0.999	0.707	0.811	0.890	0.998	0.998	0.999
gpt	0.019	0.035	0.063	0.016	0.029	0.061	0.017	0.033	0.063	0.016	0.031	0.056
Model D ($\lambda_2 = 0.5$, $\sigma_2 = 1.1$)												
pt	0.035	0.071	0.159	0.535	0.615	0.703	0.048	0.085	0.151	0.624	0.701	0.780
gpt	0.085	0.126	0.193	0.129	0.175	0.254	0.118	0.170	0.242	0.130	0.183	0.264
Model D ($\lambda_2 = 0.5$, $\sigma_2 = 1.2$)												
pt	0.243	0.364	0.509	0.997	0.998	0.999	0.205	0.296	0.432	0.999	0.999	1.000
gpt	0.312	0.417	0.538	0.499	0.599	0.714	0.448	0.545	0.670	0.535	0.632	0.739

Notes: The table reports the rejection probabilities across $1,000$ simulations for the bootstrap adjustments for the statistics defined in Equations (21)-(22), where *pt* and *gpt* follow the convention of the superscripts in the equations. ϕ denotes the density used to compute the CM statistic. Bold font indicates that the model considered satisfies the null hypothesis for the statistic in question. Models A-D are defined in Table 1.

Table 6: Simulation Results on the Choice of Density in the CM Statistic (Demeaned Y_{i1} and Y_{i2}): $n = 2000$, $T = K = 2$

ϕ	N(0,1)			N(0,3)			U(-2,2)			U(-5,5)		
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												
pt	0.027	0.048	0.088	0.032	0.051	0.086	0.029	0.055	0.094	0.031	0.052	0.084
gpt	0.023	0.046	0.082	0.029	0.045	0.082	0.029	0.046	0.078	0.029	0.044	0.080
Model B ($\lambda_2 = 0.25$)												
pt	0.027	0.048	0.085	0.033	0.053	0.085	0.030	0.049	0.090	0.030	0.049	0.081
gpt	0.022	0.042	0.079	0.028	0.046	0.077	0.028	0.049	0.080	0.029	0.044	0.078
Model B ($\lambda_2 = 0.5$)												
pt	0.022	0.048	0.080	0.030	0.051	0.091	0.030	0.053	0.091	0.029	0.049	0.084
gpt	0.023	0.042	0.082	0.028	0.044	0.082	0.028	0.047	0.082	0.027	0.045	0.080
Model C ($\lambda_2 = 0.25$)												
pt	0.922	0.944	0.978	0.997	0.998	1.000	0.888	0.928	0.954	0.998	0.998	1.000
gpt	0.025	0.044	0.083	0.028	0.043	0.082	0.026	0.045	0.078	0.027	0.044	0.081
Model C ($\lambda_2 = 0.5$)												
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	0.021	0.045	0.082	0.028	0.041	0.081	0.028	0.046	0.079	0.029	0.041	0.080
Model D ($\lambda_2 = 0.5$, $\sigma_2 = 1.1$)												
pt	0.493	0.631	0.760	0.999	1.000	1.000	0.413	0.518	0.635	1.000	1.000	1.000
gpt	0.392	0.486	0.600	0.564	0.666	0.773	0.521	0.613	0.730	0.594	0.690	0.796
Model D ($\lambda_2 = 0.5$, $\sigma_2 = 1.2$)												
pt	0.996	0.999	1.000	1.000	1.000	1.000	0.928	0.959	0.978	1.000	1.000	1.000
gpt	0.973	0.983	0.993	0.993	0.996	0.999	0.990	0.993	0.996	0.994	0.999	0.999

Notes: The table reports the rejection probabilities across $1,000$ simulations for the bootstrap adjustments for the statistics defined in Equations (21)-(22), where *pt* and *gpt* follow the convention of the superscripts in the equations. ϕ denotes the density used to compute the CM statistic. Bold font indicates that the model considered satisfies the null hypothesis for the statistic in question. Models A-D are defined in Table 1.

Table 7: Aggregated vs. Disaggregated Statistics: $n = 500$, $T = 2$, $K = 2$

	KS						CM					
	Aggregated			Disaggregated			Aggregated			Disaggregated		
	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	0.034	0.061	0.106	0.033	0.058	0.121	0.030	0.049	0.105	0.029	0.050	0.110
pt	0.013	0.023	0.052	0.011	0.025	0.050	0.023	0.048	0.092	0.024	0.041	0.090
gpt	0.005	0.013	0.028	0.003	0.008	0.017	0.013	0.031	0.076	0.017	0.026	0.062
Model B ($\lambda_2 = 0.25$)												
nt	0.361	0.489	0.630	0.485	0.593	0.714	0.351	0.468	0.593	0.347	0.457	0.593
pt	0.013	0.021	0.053	0.011	0.023	0.051	0.023	0.045	0.090	0.024	0.039	0.089
gpt	0.005	0.012	0.028	0.003	0.008	0.015	0.017	0.033	0.074	0.016	0.025	0.062
Model B ($\lambda_2 = 0.5$)												
nt	0.937	0.973	0.989	0.985	0.992	0.997	0.947	0.967	0.987	0.943	0.966	0.983
pt	0.014	0.023	0.054	0.012	0.023	0.050	0.023	0.044	0.089	0.025	0.041	0.088
gpt	0.004	0.012	0.028	0.003	0.008	0.016	0.016	0.032	0.075	0.015	0.025	0.060
Model C ($\lambda_2 = 0.25$)												
nt	0.308	0.423	0.598	0.465	0.562	0.682	0.275	0.366	0.503	0.289	0.387	0.522
pt	0.340	0.442	0.585	0.403	0.534	0.637	0.339	0.439	0.558	0.349	0.459	0.587
gpt	0.002	0.011	0.028	0.004	0.008	0.017	0.017	0.030	0.069	0.017	0.030	0.055
Model C ($\lambda_2 = 0.5$)												
nt	0.929	0.967	0.989	0.988	0.992	0.996	0.900	0.944	0.974	0.921	0.953	0.978
pt	0.965	0.982	0.993	0.984	0.993	0.996	0.945	0.965	0.982	0.950	0.969	0.984
gpt	0.001	0.012	0.028	0.004	0.009	0.019	0.017	0.028	0.071	0.014	0.027	0.055
Model D ($\lambda_2 = 0.5$, $\sigma_2 = 1.1$)												
nt	0.985	0.995	1.000	0.992	0.997	0.998	0.934	0.963	0.985	0.928	0.959	0.984
pt	0.411	0.523	0.662	0.505	0.591	0.706	0.105	0.162	0.259	0.104	0.155	0.265
gpt	0.009	0.027	0.059	0.013	0.024	0.046	0.070	0.120	0.198	0.095	0.153	0.245
Model D ($\lambda_2 = 0.5$, $\sigma_2 = 1.2$)												
nt	1.000	1.000	1.000	1.000	1.000	1.000	0.944	0.970	0.992	0.951	0.974	0.991
pt	0.972	0.988	0.995	0.989	0.995	0.998	0.394	0.513	0.666	0.389	0.517	0.660
gpt	0.053	0.100	0.181	0.068	0.120	0.209	0.253	0.343	0.458	0.429	0.545	0.650

Notes: The table reports rejection probabilities across 1,000 simulations for the bootstrap adjustments for the aggregated statistics defined in Equations (13)-(15) and (17)-(19) as well as the disaggregated statistics defined in Equations (23)-(25) and (26)-(28). 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 8: Aggregated vs. Disaggregated Statistics: $n = 500$, $T = 2$, $K = 16$

	KS						CM					
	Aggregated			Disaggregated			Aggregated			Disaggregated		
	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	0.017	0.032	0.074	0.127	0.202	0.319	0.015	0.040	0.106	0.138	0.186	0.280
pt	0.002	0.009	0.023	0.003	0.005	0.008	0.008	0.019	0.043	0.039	0.078	0.129
gpt	0.002	0.005	0.011	0.000	0.000	0.001	0.003	0.014	0.030	0.000	0.003	0.015
Model B ($\lambda_2 = 0.25$)												
nt	0.297	0.434	0.601	0.645	0.763	0.862	0.670	0.757	0.845	0.604	0.721	0.845
pt	0.002	0.011	0.024	0.004	0.005	0.009	0.011	0.018	0.050	0.043	0.076	0.138
gpt	0.002	0.003	0.010	0.000	0.000	0.001	0.004	0.013	0.028	0.000	0.002	0.013
Model B ($\lambda_2 = 0.5$)												
nt	0.892	0.942	0.980	0.991	0.998	1.000	0.993	0.995	0.997	0.992	0.996	0.999
pt	0.003	0.011	0.024	0.004	0.005	0.008	0.009	0.018	0.045	0.037	0.078	0.131
gpt	0.002	0.005	0.013	0.000	0.000	0.001	0.005	0.012	0.031	0.000	0.002	0.017
Model C ($\lambda_2 = 0.25$)												
nt	0.198	0.311	0.479	0.642	0.754	0.866	0.467	0.585	0.718	0.590	0.713	0.844
pt	0.098	0.167	0.280	0.016	0.047	0.102	0.321	0.442	0.566	0.282	0.420	0.599
gpt	0.002	0.004	0.010	0.000	0.000	0.001	0.004	0.016	0.027	0.000	0.003	0.011
Model C ($\lambda_2 = 0.5$)												
nt	0.801	0.902	0.959	0.988	0.994	0.999	0.975	0.991	0.998	0.995	1.000	1.000
pt	0.631	0.734	0.883	0.500	0.657	0.805	0.954	0.969	0.986	0.960	0.989	0.997
gpt	0.002	0.004	0.012	0.000	0.000	0.001	0.006	0.013	0.030	0.000	0.002	0.012
Model D ($\lambda_2 = 0.5$, $\sigma_2 = 1.1$)												
nt	0.817	0.914	0.967	0.993	0.997	1.000	0.990	0.997	0.998	0.993	0.999	1.000
pt	0.014	0.027	0.064	0.003	0.007	0.019	0.024	0.044	0.094	0.047	0.087	0.157
gpt	0.003	0.006	0.012	0.000	0.000	0.000	0.004	0.014	0.032	0.000	0.002	0.018
Model D ($\lambda_2 = 0.5$, $\sigma_2 = 1.2$)												
nt	0.769	0.884	0.951	0.987	0.999	1.000	0.990	0.997	0.999	0.990	0.997	1.000
pt	0.057	0.112	0.205	0.015	0.037	0.081	0.085	0.150	0.254	0.077	0.135	0.245
gpt	0.003	0.006	0.012	0.000	0.001	0.002	0.003	0.017	0.034	0.008	0.019	0.053

Notes: The table reports rejection probabilities across 1,000 simulations for the bootstrap adjustments for the aggregated statistics defined in Equations (13)-(15) and (17)-(19) as well as the disaggregated statistics defined in Equations (23)-(25) and (26)-(28). 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 9: Aggregated vs. Disaggregated Statistics: $n = 2000$, $T = 2$ $K = 2$

	KS						CM					
	Aggregated			Disaggregated			Aggregated			Disaggregated		
	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	0.030	0.063	0.099	0.037	0.062	0.115	0.029	0.053	0.109	0.029	0.049	0.108
pt	0.022	0.043	0.072	0.023	0.039	0.073	0.033	0.067	0.113	0.029	0.052	0.102
gpt	0.008	0.020	0.050	0.007	0.016	0.039	0.034	0.054	0.101	0.024	0.048	0.088
Model B ($\lambda_2 = 0.25$)												
nt	0.930	0.973	0.989	0.985	0.991	0.996	0.947	0.974	0.993	0.950	0.973	0.991
pt	0.022	0.041	0.077	0.026	0.039	0.072	0.034	0.062	0.109	0.030	0.053	0.100
gpt	0.011	0.023	0.052	0.007	0.017	0.036	0.032	0.050	0.102	0.022	0.042	0.088
DGP2 ($\lambda_2 = 0.5$)												
nt	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
pt	0.022	0.043	0.071	0.024	0.039	0.076	0.033	0.067	0.113	0.029	0.051	0.100
gpt	0.009	0.022	0.050	0.007	0.017	0.037	0.033	0.053	0.101	0.026	0.045	0.088
Model C ($\lambda_2 = 0.25$)												
nt	0.936	0.967	0.987	0.987	0.993	0.995	0.934	0.963	0.983	0.942	0.964	0.986
pt	0.973	0.988	0.993	0.989	0.993	0.998	0.965	0.979	0.995	0.969	0.982	0.995
gpt	0.008	0.021	0.049	0.007	0.015	0.036	0.026	0.051	0.101	0.020	0.041	0.087
Model C ($\lambda_2 = 0.5$)												
nt	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
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	0.009	0.022	0.049	0.005	0.015	0.034	0.033	0.054	0.099	0.022	0.045	0.089
Model D ($\lambda_2 = 0.5$, $\sigma_2 = 1.1$)												
nt	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
pt	0.988	0.993	0.998	0.993	0.996	1.000	0.556	0.681	0.804	0.552	0.689	0.807
gpt	0.108	0.168	0.278	0.132	0.201	0.309	0.321	0.423	0.554	0.472	0.573	0.687
Model D ($\lambda_2 = 0.5$, $\sigma_2 = 1.2$)												
nt	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
pt	1.000	1.000	1.000	1.000	1.000	1.000	0.995	0.999	1.000	0.996	0.998	1.000
gpt	0.562	0.714	0.835	0.774	0.867	0.923	0.910	0.949	0.982	0.994	0.997	0.999

Notes: The table reports rejection probabilities across 1,000 simulations for the bootstrap adjustments for the aggregated statistics defined in Equations (13)-(15) and (17)-(19) as well as the disaggregated statistics defined in Equations (23)-(25) and (26)-(28). 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 10: Aggregated vs. Disaggregated Statistics: $n = 2000$, $T = 2$, $K = 16$

	KS						CM					
	Aggregated			Disaggregated			Aggregated			Disaggregated		
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	0.026	0.048	0.088	0.070	0.120	0.206	0.031	0.052	0.096	0.126	0.174	0.252
pt	0.007	0.014	0.032	0.004	0.006	0.018	0.021	0.034	0.073	0.087	0.128	0.183
gpt	0.008	0.016	0.030	0.000	0.000	0.002	0.015	0.026	0.052	0.001	0.007	0.019
Model B ($\lambda_2 = 0.25$)												
nt	0.968	0.987	0.998	0.996	0.997	1.000	1.000	1.000	1.000	0.978	0.992	0.996
pt	0.008	0.015	0.036	0.002	0.006	0.015	0.024	0.035	0.075	0.088	0.126	0.185
gpt	0.009	0.016	0.033	0.000	0.000	0.002	0.018	0.034	0.053	0.004	0.008	0.025
Model B ($\lambda_2 = 0.5$)												
nt	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
pt	0.008	0.016	0.037	0.002	0.006	0.016	0.022	0.038	0.075	0.087	0.130	0.187
gpt	0.010	0.016	0.033	0.000	0.000	0.002	0.018	0.034	0.054	0.003	0.009	0.026
Model C ($\lambda_2 = 0.25$)												
nt	0.908	0.951	0.985	0.997	1.000	1.000	0.993	0.998	0.999	0.977	0.993	0.997
pt	0.811	0.891	0.945	0.721	0.847	0.936	0.987	0.995	0.996	0.965	0.984	0.993
gpt	0.007	0.016	0.033	0.000	0.000	0.001	0.016	0.027	0.057	0.002	0.007	0.024
Model C ($\lambda_2 = 0.5$)												
nt	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
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	0.010	0.017	0.030	0.000	0.000	0.001	0.013	0.032	0.054	0.003	0.006	0.029
Model D ($\lambda_2 = 0.5$, $\sigma_2 = 1.1$)												
nt	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
pt	0.143	0.222	0.331	0.053	0.119	0.206	0.155	0.235	0.378	0.138	0.206	0.305
gpt	0.010	0.016	0.029	0.001	0.001	0.004	0.016	0.034	0.067	0.024	0.052	0.098
Model D ($\lambda_2 = 0.5$, $\sigma_2 = 1.2$)												
nt	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
pt	0.684	0.790	0.899	0.678	0.793	0.896	0.692	0.808	0.906	0.444	0.605	0.798
gpt	0.009	0.019	0.038	0.000	0.006	0.033	0.024	0.041	0.077	0.168	0.254	0.362

Notes: The table reports rejection probabilities across 1,000 simulations for the bootstrap adjustments for the aggregated statistics defined in Equations (13)-(15) and (17)-(19) as well as the disaggregated statistics defined in Equations (23)-(25) and (26)-(28). 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 11: 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 3.3.4. $std(.)$ denotes the standard deviation of the variable in the brackets.

Table 12: 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 ($\lambda_{t+1} - \lambda_t = 0.01$)												
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 B ($\lambda_{t+1} - \lambda_t = 0.05$)												
pt	0.007	0.018	0.045	0.013	0.025	0.056	0.012	0.027	0.059	0.015	0.030	0.057
gpt	0.005	0.008	0.015	0.004	0.008	0.017	0.003	0.009	0.026	0.004	0.009	0.017
Model C ($\lambda_{t+1} - \lambda_t = 0.01$)												
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 C ($\lambda_{t+1} - \lambda_t = 0.05$)												
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	0.004	0.006	0.015	0.003	0.008	0.018	0.007	0.012	0.025	0.002	0.009	0.017
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 for the bootstrap adjustments for the statistics defined in Equations (29)-(32). 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 13: Simulation Resembling Subsample of NLSY 1983-1987 (CM with Zero-Mean Density): $n = 1000$, $T = 5$, $K = 16$

	CM								
ϕ	N(0,0.25)			N(0,0.5)			U(-7,7)		
α	0.025	0.05	0.10	0.025	0.05	0.10	0.025	0.05	0.10
Model A									
pt	0.001	0.006	0.011	0.001	0.005	0.017	0.001	0.004	0.013
gpt	0.001	0.005	0.019	0.002	0.005	0.017	0.002	0.005	0.017
Model B ($\lambda_{t+1} - \lambda_t = 0.01$)									
pt	0.002	0.005	0.011	0.002	0.006	0.017	0.001	0.003	0.012
gpt	0.002	0.006	0.020	0.003	0.005	0.018	0.004	0.006	0.018
Model B ($\lambda_{t+1} - \lambda_t = 0.025$)									
pt	0.001	0.006	0.012	0.000	0.006	0.017	0.001	0.003	0.015
gpt	0.003	0.006	0.020	0.003	0.006	0.021	0.002	0.006	0.017
Model B ($\lambda_{t+1} - \lambda_t = 0.05$)									
pt	0.003	0.005	0.012	0.002	0.006	0.016	0.002	0.006	0.013
gpt	0.002	0.007	0.016	0.004	0.006	0.016	0.002	0.006	0.015
Model C ($\lambda_{t+1} - \lambda_t = 0.01$)									
pt	0.192	0.329	0.538	0.171	0.280	0.499	0.138	0.237	0.422
gpt	0.003	0.007	0.019	0.004	0.008	0.019	0.002	0.007	0.017
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
gpt	0.002	0.007	0.017	0.003	0.007	0.016	0.003	0.006	0.016
Model C ($\lambda_{t+1} - \lambda_t = 0.05$)									
pt	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
gpt	0.002	0.006	0.016	0.001	0.006	0.018	0.002	0.005	0.014
Model D ($\lvert\sigma_{t+1} - \sigma_t\rvert = 0.025$)									
pt	0.098	0.150	0.243	0.052	0.084	0.152	0.114	0.175	0.279
gpt	0.009	0.014	0.051	0.007	0.018	0.054	0.005	0.017	0.051
Model D ($\lvert\sigma_{t+1} - \sigma_t\rvert = 0.05$)									
pt	0.846	0.902	0.957	0.628	0.718	0.821	0.884	0.929	0.969
gpt	0.108	0.152	0.220	0.092	0.139	0.215	0.105	0.157	0.222
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
gpt	0.834	0.900	0.948	0.814	0.879	0.932	0.844	0.906	0.949

Notes: The table reports rejection probabilities across 1,000 simulations for the bootstrap adjustments for the CM statistic defined in Equation (33)-(34) using the densities in the column title. Bold font indicates that the model considered satisfies the null hypothesis for the statistic in question. Models A-D are defined in Table 1.

Table 14: 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)
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

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. $P_n(S_{i,t-1} = S_{it})$ denotes the empirical probability of the event in brackets.

Table 15: Returns to Schooling: ANACOVA Results

	Full Sample			1983-1984		
	(1)	(2)	(3)	(1)	(2)	(3)
S	-0.0688	0.0772	0.0714	-0.2343	-0.0196	-0.0097
	(0.1308)	(0.0152)	(0.0148)	(0.1502)	(0.0194)	(0.0184)
S^2	-0.0070	-0.0005		-0.0013	0.0006	
	(0.0048)	(0.0004)		(0.0060)	(0.0005)	
Age^2	-0.0030		-0.0001	-0.0023		0.0006
	(0.0005)		(0.0003)	(0.0009)		(0.0004)
$S * Age$	0.0134			0.0113		
	(0.0015)			(0.0032)		
$Union$	0.1397	0.1423	0.1424	0.1642	0.1642	0.1643
	(0.0162)	(0.0163)	(0.0163)	(0.0152)	(0.0152)	(0.0152)

Notes: The above results are replication of the ANACOVA results in Angrist and Newey (1991, Table 3) for the full-sample as well as 1983-1984.

Table 16: 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
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

Notes: The above reports p-values of the tests for time homogeneity using the NLSY subsample, $n = 1,087$, where schooling S_{it} is the sole regressor. 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 (40).

Table 17: APE of Schooling on Log Hourly Wage: NLSY 1983-1987

S		Subs	APE	S.E.	t-Stat
1983	1984				
11	12	26	0.052	0.089	0.576
12	13	24	-0.120	0.133	-0.904
13	14	28	-0.069	0.061	-1.124
14	15	19	-0.037	0.035	-1.045
15	16	20	0.141	0.131	1.073
All Movers		122	-0.012	0.043	-0.267
1984	1985				
11	12	6	-0.010	0.032	-0.318
12	13	14	-0.080	0.067	-1.191
13	14	12	0.295	0.247	1.194
14	15	17	0.011	0.063	0.178
15	16	17	0.263	0.111	2.379
All Movers		73	0.095	0.055	1.723
1985	1986				
11	12	3	0.457	0.278	1.644
12	13	12	0.148	0.113	1.311
13	14	11	0.206	0.170	1.214
14	15	14	0.052	0.106	0.496
15	16	11	0.601	0.112	5.345
All Movers		58	0.226	0.060	3.737
1986	1987				
13	14	9	0.030	0.119	0.250
14	15	8	-0.125	0.168	-0.743
15	16	11	0.167	0.099	1.695
All Movers		41	-0.012	0.068	-0.179

Note: The APE for all movers is calculated using (42). All other APEs are computed using (41). *Subs* denotes subsample size. APEs are reported for subsamples with $Subs \geq 5$.

Table 18: Testing Time Homogeneity with a Parallel Trend of Log Earnings ($X_{it} = (S_{it}, Union_{it})'$)

Statistic ϕ	KS	CM $N(6.5, 0.25)$	CM $N(6.5, 0.5)$	F
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 p-values of the tests for time homogeneity using the NLSY subsample, $n = 1,087$, where S_{it} and $Union_{it}$ are regressors. 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 (40), where S_{it} is replaced by both S_{it} and $Union_{it}$.

Table 19: APE of Schoooling on Log Hourly Wage (*Union Status Fixed*): NLSY 1983-1987

S		*Union*	Subs	APE	S.E.	t-Stat
1983	1984	1983/84				
11	12	0	19	0.067	0.078	0.863
12	13	0	19	-0.062	0.122	-0.509
13	14	0	23	-0.081	0.055	-1.475
14	15	0	14	-0.021	0.031	-0.672
15	16	0	20	0.137	0.131	1.045
All Movers			108	-0.001	0.040	-0.014
1984	1985	1984/85				
12	13	0	8	-0.104	0.047	-2.232
13	14	0	9	0.177	0.252	0.704
14	15	0	13	0.014	0.035	0.393
15	16	0	13	0.313	0.119	2.630
16	17	0	5	0.050	0.102	0.488
All Movers			61	0.075	0.051	1.468
1985	1986	1985/86				
12	13	0	7	-0.014	0.146	-0.092
13	14	0	7	0.158	0.230	0.689
14	15	0	8	-0.050	0.059	-0.847
15	16	0	7	0.520	0.136	3.824
All Movers			44	0.144	0.061	2.363
1986	1987	1986/87				
13	14	0	8	0.101	0.112	0.904
14	15	0	7	-0.111	0.194	-0.574
15	16	0	7	0.231	0.122	1.890
All Movers			33	0.018	0.073	0.248