Testing for Distributional Treatment Effects: A Set Identification Approach

Sung Jae Jun*, Yoonseok Lee†, and Youngki Shin‡

April 2011

Abstract

We propose testing procedure for stochastic dominance between potential outcomes in the panel treatment setup. Unobserved heterogeneity in the model need not be time–invariant and is allowed to be arbitrarily correlated with a treatment variable. Instead of imposing additional restrictions such as unconfoundedness or the availability of instruments, we take a set identification approach. We treat the identified bounds as model restrictions and test if these restrictions are compatible with stochastic dominance between counterfactual outcomes. Extensions to higher order dominance and heterogeneous effects given covariates are also considered. By applying the proposed methods we investigate distributional effects of smoking during pregnancy on infant's birthweight.

JEL Classification: C12, C21, C23.

Keywords: treatment effects, stochastic dominance, set identification, unobserved heterogeneity, panel data.

*Department of Economics, Pennsylvania State University, 619 Kern Bldg, University Park, PA 16802, sjun@psu.edu

†Department of Economics, University of Michigan, 611 Tappan St., Ann Arbor, MI 48109, yoolee@umich.edu

‡(Corresponding author) Department of Economics, University of Western Ontario, 1151 Richmond St., London, ON N6A 5C2, Canada, yshin29@uwo.ca

1. Introduction

We propose testing procedure for stochastic dominance between potential outcomes in the panel treatment setup. Unobserved heterogeneity in the model need not be time–invariant and is allowed to be arbitrarily correlated with a treatment variable. Instead of imposing additional restrictions such as unconfoundedness or the availability of instruments, we take a set identification approach. We treat the identified bounds as model restrictions and test if these restrictions are compatible with stochastic dominance between counterfactual outcomes. Extensions to higher order dominance and heterogeneous effects given covariates are also considered. By applying the proposed methods we investigate distributional effects of smoking during pregnancy on infant's birthweight.

Evaluating policy or treatment effects has been an important question in economics and other fields of the social sciences. Unlike the natural sciences, it is a common concern in the social sciencses that a treatment variable may be endogenous due to uncontrolled heterogeneity and/or self-selection.[1] As well summarized by Imbens and Wooldridge (2009) and Heckman and Vytlacil (2007a,b), various econometric methods have been developed to address this issue, and a myriad of empirical studies were conducted based on them. They can be categorized broadly into three groups: methods based on the assumption of unconfoundedness and matching (e.g., Dehejia and Wahba (2002)), methods based on instrumental variables (e.g., Angrist, Imbens, and Rubin (1996)), and methods based on panel data using the idea of difference-in-differences (DID; e.g., Ashenfelter and Card (1985)). However, each approach still faces some potential problems in certain situations. Unobserved heterogeneity may not be well-described by the observed characteristics; good instrumental variables may be hard to find; and the additive separability of unobserved heterogeneity required by the DID approach can be restrictive.

In addition, many of the existing methods focus on average treatment effects while there is accumulating evidence that heterogeneity in treatment effects matters and possibly misleads the researcher who just focuses on the average effect. For instance, Heckman and Smith (1997) considered program evaluation, and they provided empirical evidence that heterogeneity in program responses is indeed important. Also, Djebbari and Smith (2008) provided extra evidence for impact heterogeneity using quasi-experimental data from Mexico.

Some recent research have investigated distributional treatment effects under various assumptions. Firpo (2007) used the unconfoundedness assumption to identify quantile treatment effects. Lee (2009) proposed nonparametric tests for the lack of distributional effects, where he followed the two-sample setup of the Mann–Whitney test. Crump, Hotz,

[1]Some empirical studies are based on quasi-experiments, where randomization is imposed by design. This is not the situation we consider. We focus on the case where data are observational.

Imbens, and Mitnik (2008) and Lee and Whang (2009) proposed testing methods for heterogeneous treatment effects under the unconfoundedness assumption. Abadie (2002) proposed bootstrap tests for stochastic dominance between potential outcomes given instrumental variables. Rothe (2010) discussed treatment effects when the counterfactual experiment is to change the distribution of a treatment variable. We note that all these approaches depend on the conventional assumptions—either unconfoundedness or the availability of instruments—that may not hold in some situations as illustrated below.

All these points above necessitate studying distributional treatment effects with a minimal set of assumptions. The first question we cast in this paper is how much we can learn about distributional treatment effects without making conventional assumptions. Specifically, we consider a model where unobserved heterogeneity is arbitrarily correlated with a treatment variable but panel data are available.

This question is further motivated by the fact that there are many examples where the endogeneity issue is quite severe but good panel data sets are available. We mention three examples here. The first one is the study on the wage premium of the union membership. The union membership status is usually believed to be correlated with unobserved ability. This subject has been investigated by many labor economists using various panel data sets: the National Longitudinal Survey (NLS), the Current Population Survey (CPS), the Canadian Labor Market Activity Survey (LMAS), and the German Socio-economic Panel (see, e.g., Jones (1982), Blakemore, Hunt, and Kiker (1986), Robinson (1989), Lemieux (1998), Budd and Na (2000), and Beck and Fitzenberger (2004)).

The second example of repeated treatments can be found in Health Maintenance Organization (HMO) markets. For instance, Jung (2010) examined the effect of voluntary information disclosure of HMO plans on service quality using panel data. Information released to the public may help consumers recognize quality differences among different providers, in which case it becomes a policy-relevant question how much public information disclosure will improve service quality of each plan.

Finally, in the health economics literature, we can find many studies that measure the effect of smoking during pregnancy on birth outcomes (such as birghweight). In the panel setup, the (potential) outcome variable is the weight of the t^{th} baby delivered by mother i, and the treatment variable is the smoking status of the mother during the pregnancy of each baby. The usual concern is that the smoking behavior is correlated with other health-related factors such as nutrition status or drug addiction that are rarely observed. To deal with such an endogeneity issue, Abrevaya (2006) carefully constructed a pseudo-panel data set based on the U.S. federal natality data, and used a fixed-effect estimator.

It should be noted that all the empirical studies listed above use a linear panel model, which is a natural starting point. The fixed-effect estimator, in fact, can handle unobserved heterogeneity effectively when heterogeneity is *time-invariant* and *additively separable*. However, if either one of these assumptions is violated, then the results will be questionable. For example, a woman having stopped smoking at her second pregnancy may also have changed some health-related (unobserved) factors. In this case the assumption of time-invariant heterogeneity is questionable. In the union premium example, workers with low ability may have different treatment effects from those with high ability, in which case additive separability of unobserved ability may be unrealistic. One way of relaxing these restrictions is to introduce correlated random coefficients while the basic linear structure is maintained (see, e.g., Wooldridge (2005), Graham and Powell (2008), and references therein). Another approach is to remove the linearity assumption and to work with a nonparametric structural function including a nonseparable unobserved heterogeneity term (e.g., Chernozhukov, Fernandez-Val, and Newey (2009)). Depending on the level of nonparametric restrictions, we can think of various nonseparable models as listed in Matzkin (2007). In this paper we follow the latter approach but incorporate time-varying unobserved heterogeneity to capture distributional treatment effects with minimal assumptions.

The second question is about practical inference methods based on set identification. As a result of flexible model restrictions, the distribution functions of potential outcomes, the objects of interest in our setup, are only partially identified. Inference under partial identification has been actively studied during the last decade, and there are many methods available: e.g., Chernozhukov, Hong, and Tamer (2007), Beresteanu and Molinari (2008), Rosen (2008), Romano and Shaikh (2010), Andrews and Soares (2010), just to name a few. Although many papers, once partial identification is obtained, refer to these studies for possible inference, there is only a little empirical work that actually conducts set inferences with real data. We think of two reasons for the gap between the theoretical development and its empirical impact. The first reason is the cost of flexibility: The size of a confidence region is often too large to draw any general conclusion from data. Second, asymptotic theory in this area is quite complicated, which reduces the accessibility of empirical researchers. We emphasize here that there are potentially many policy-relevant questions that can be empirically answered without conducting a complicated set inference. In particular, by focusing on a particular question of interest, we can often find much simpler econometric procedures with more accessible distribution theory. The recent idea proposed by Hahn and Ridder (2009) highlights the point. They focused on an inference on one particular parameter of interest while the identified set (of the entire parameters) is treated as a model restriction that should be maintained both under the null and under the alternative. Using a similar idea, we show that stochastic dominance of the distributions of potential outcomes can be tested without conducting a complicated set inference. Furthermore, we apply our method to real data; we analyze the panel data on infant's birthweight used by Abrevaya (2006).

Contributions of this paper can be summarized as follows. First, we propose a method for testing distributional treatment effects in a flexible setup. Specifically, we use panel data without imposing the standard assumptions such as unconfoundedness, time-invariance and/or additive separability of unobserved heterogeneity, and the availability instruments. Second, we bridge the gap between the theoretical development in the set inference literature and its application to empirical studies in the treatment effect context. Finally, empirical findings about the causal effects of smoking on birth outcomes add more rigorous scientific evidence to the literature.

The remainder of the paper is organized as follows. In Section 2 we introduce the basic framework and discuss hypotheses of interest based on the partial identification results. In Section 3 we propose test statistics and establish their asymptotic properties. In Section 4 two extensions of the approach are discussed. Section 5 presents an empirical study that investigates the effect of smoking on birthweight. Section 6 concludes with some remarks.

2. The Framework

In this section we discuss the basic framework and partial identification. We then show how to formulate the hypotheses of stochastic dominance by using the partial identification result.

2.1. **The Parameters and Their Bounds.**

We consider the panel data $\{(Y_{it}, D_{it}) : i = 1, 2, \cdots, n, t = 1, 2, \cdots, T\}$, where D_{it} is a binary treatment variable and Y_{it} is an outcome variable of interest. We assume a short panel, where n is large but T is small and fixed. This also corresponds to the examples discussed in the introduction. The observed outcome Y_{it} depends on the treatment D_{it} in the following way:

$$Y_{it} = D_{it}Y_{it}^1 + (1 - D_{it})Y_{it}^0,$$

where Y_{it}^1 and Y_{it}^0 are potential outcomes when $D_{it} = 1$ and 0, respectively. We observe only one of these potential outcomes depending on the treatment status D_{it}.

The counterfactual setup like this is now standard, at least in the cross-section context, where the common objective is to compare (some features of) the distributions of two potential outcomes; e.g., Rosenbaum and Rubin (1983), Hahn (1998), Hirano, Imbens, and Ridder (2003), and Firpo (2007). This objective is usually achieved by assuming randomized treatment conditional on observed heterogeneity. However, as we emphasized in the introduction, it is one of our main goals to avoid this standard assumption to allow for potential endogeneity. Data on repeated treatment turn out to be useful for this purpose, as we discuss below.

We first introduce some notation. Let A_{it} represent a vector of heterogeneity, some components of which are not observed, and let $A_i = (A'_{i1}, \cdots, A'_{iT})'$ be the history of the heterogeneity of person i. For the sake of presentational clarity, we will not explicitly use observed covariates for now, but we will discuss such extension in Section 4.2. Note that each component of the unobserved heterogeneity A_i may vary over time. Let $\mu_{A_i}(\cdot)$ and $\mu_{A_i|B_i}(\cdot|b)$ be the marginal and conditional distribution functions of A_i given a generic random variable $B_i = b$. Further, letting $F_t^j(\cdot)$ be the distribution function of Y_{it}^j for $j = 0, 1$, we can write

$$F_t^j(y) = \int \mathbb{P}\left[Y_{it}^j \leq y \middle| A_i = a\right] d\mu_{A_i}(a). \tag{1}$$

We now make two assumptions on the underlying model.

Assumption 1 (Selection on Unobservables)**.** *The treatment history $D_i = (D_{i1}, \cdots, D_{iT})'$ is independent of the current potential outcomes (Y_{it}^0, Y_{it}^1) conditional on the history of all heterogeneity A_i.*

Assumption 2 (Time Homogeneity)**.** *For all $t \neq s$ and $j = 0, 1$, the conditional distribution of Y_{it}^j is the same as that of Y_{is}^j given the history of all heterogeneity A_i.*

Assumption 1 is more flexible than the standard unconfoundedness assumption since the heterogeneity, A_{it}, which can be arbitrarily correlated with the treatment, does not need to be observed. One practical restriction imposed by Assumption 1 is that there is no feedback from Y_{it-1} to D_{it} once A_i is controlled for. This requirement can be viewed as the strong exogeneity assumption given A_i. We can relax it to the assumption with predetermined treatments: (Y_{it}^0, Y_{it}^1) is independent of $(D_{i1}, \cdots, D_{it})'$ given A_i but can be correlated with D_{is} for $s > t$. However, we do not discuss this extension in this paper, because it does not change the essence of the analysis but makes presentation complicated.

Assumption 2 means that the distributions of potential outcomes given the history of heterogeneity are stationary. Note that Y_{it}^j and Y_{is}^j, however, can be still correlated with each other. This assumption implies that $F_t^j(\cdot)$ in (1) does not depend on t anymore, and we will simply denote it by $F^j(\cdot)$ throughout the paper. A similar assumption is also adopted by Chernozhukov, Fernandez-Val, Hahn, and Newey (2009) and Khan, Ponomareva, and Tamer (2011) in the context of a nonseparable panel regression model and a censored panel regression model, respectively.

We now discuss the identification of $F^j(y)$, focusing on $j = 1$. Note first that

$$F^1(y) = \mathbb{P}\left[D_{it} = 0\right] \int \mathbb{P}\left[Y_{it}^1 \leq y \middle| A_i = a\right] d\mu_{A_i|D_{it}}(a|0) + \mathbb{P}\left[Y_{it} \leq y,\ D_{it} = 1\right], \tag{2}$$

because the second term on the right-hand side is equal to

$$\mathbb{P}\left[D_{it}=1\right]\int\mathbb{P}\left[Y_{it}^{1}\leq y\middle|A_{i}=a\right]d\mu_{A_{i}|D_{it}}(a|1)$$

from Assumption 1. The first term in (2) is not identified because Y_{it}^1 is not observed when $D_{it}=0$. However, it is trivially bounded between 0 and $\mathbb{P}[D_{it}=0]$, which results in the interval identification of $F^1(y)$. Lemma 1 below shows that this interval can be further improved by using the observations of multiple periods.

To this end, we need some additional notation. Let $(t_1,t_2,\cdots,t_T)$ be a permutation of $(1,2,\cdots,T)$. For $j\in\{0,1\}$, let $p_1^j(y)=\mathbb{P}[Y_{it_1}\leq y,D_{it_1}=j]$ and $p_k^j(y)=\mathbb{P}[Y_{it_k}\leq y,D_{it_1}=1-j,\cdots,D_{it_{k-1}}=1-j,D_{it_k}=j]$ for $k\geq 2$. Then, define

$$\begin{aligned}L^j(y)&=\sum_{k=1}^{T}p_k^j(y),\\U^j(y)&=L^j(y)+\mathbb{P}[D_{it_1}=1-j,D_{it_2}=1-j,\cdots,D_{it_T}=1-j].\end{aligned}$$

The bounds of the distributions $F^j(\cdot)$ are provided in the following lemma.

Lemma 1. *Suppose that Assumptions 1 and 2 are satisfied. For each* $j\in\{0,1\}$ *, we have*

$$0\leq L^j(y)\leq F^j(y)\leq U^j(y)\leq 1$$

for all $y\in\mathbb{R}$.

Similar calculation to Lemma 1 can be found in Manski (1990) and Chernozhukov, Fernandez-Val, and Newey (2009). Note that the dependence of each bound on a particular permutation is implicit in Lemma 1. In fact, the bounds in Lemma 1 should be the same in any permutation of $(1,2,\cdots,T)$ because of the time homogeneity assumption.

2.2. **The Hypotheses.** The testing problem considered is the stochastic dominance of Y^0 to Y^1, where the distributions of Y^0 and Y^1 are partially identified. Recall that Y^0 stochastically dominates Y^1 in the first order if and only if

$$F^0(y)\leq F^1(y)\quad\text{for all}\quad y\in\mathbb{R}. \tag{3}$$

If the null hypothesis is given as the condition in (3), it means that the treatment $D=1$ has a distributional treatment effect to the negative direction. For example, one might test whether the birthweight decreases over all quantiles when a mother smokes.[2] Note also that

[2]This interpretation requires a rather strong assumption such that each individual maintain his/her rank in both treated and untreated distributions. Because of this restriction, Heckman and Smith (1997) claimed that the quantile treatment effect be based on the joint distribution, not the difference of two marginals. See Firpo (2007) and Imbens and Wooldridge (2009) for more discussions.

this hypothesis is stronger than comparing the average treatment effects, i.e., $\mathbb{E}[Y_{it}^0] \geq \mathbb{E}[Y_{it}^1]$ is implied by (3).

Since $F^1(\cdot) - F^0(\cdot)$ is not point-identified, it is infeasible to test (3) directly. We may think of three different approaches to this problem. First, one may construct a set confidence region for $F^1(\cdot) - F^0(\cdot)$ directly. The identification result in Lemma 1 implies that for all $y \in \mathbb{R}$

$$L^1(y) - U^0(y) \leq F^1(y) - F^0(y) \leq U^1(y) - L^0(y).$$

Constructing a confidence region based on these bounds and checking if the region lies above zero would be one testing procedure. However, this method is unnecessarily conservative.

Second, the method of testing for moment inequalities can be considered. Letting $\gamma(y) = F^1(y) - F^0(y)$, we can reparameterize $(F^1(y), F^0(y))$ as $(F^1(y), \gamma(y))$. A simple algebra shows that

$$\begin{pmatrix} L^0(y) \leq F^1(y) - \gamma(y) \leq U^0(y) \\ L^1(y) \leq F^1(y) \leq U^1(y) \\ \gamma(y) \geq 0 \end{pmatrix} \Longleftrightarrow \begin{pmatrix} -1 & 1 \\ 1 & -1 \\ -1 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} F^1(y) \\ \gamma(y) \end{pmatrix} \leq \begin{pmatrix} -L^0(y) \\ U^0(y) \\ -L^1(y) \\ U^1(y) \\ 0 \end{pmatrix}.$$

Now for each $y \in \mathbb{R}$ the utmost right-hand side term is a vector of expectations, and one can view $(F^1(y), \gamma(y))$ as a parameter satisfying a set of moment inequalities. Therefore, we can use any of the existing inference procedures for moment inequality models (e.g., Rosen (2008), and Andrews and Soares (2010)), and conduct joint inferences on $(F^1(y), \gamma(y))$. Projecting it on the space of $\gamma(y)$ and checking if the (projected) region lies above zero will establish a testing procedure for stochastic dominance.

This method would improve the power compared to the first approach, but there are still some loss of power due to the inference involving $F^1(y)$. Furthermore, note that the parameter in our testing problem is an infinite-dimensional object. Although we express both procedures as a finite number of inequalities, there are in fact infinite number of them indexed by y. Constructing confidence regions in the infinite-dimensional space is not so obvious.

The third approach, which we have adopted in this paper, is to treat the identified bounds in Lemma 1 as given restrictions and to test whether stochastic dominance is compatible with the restrictions. Therefore, our approach stands in the same line as those of Imbens and Manski (2004) and Hahn and Ridder (2009) in the sense that we test the existence of the true parameter satisfying the null hypothesis of dominance as well as the restrictions given by the identified bounds. This approach is less conservative than the aforementioned ones, and the asymptotic theory involved turns out to be simple and standard, which was highlighted by Hahn and Ridder (2009) in the context of moment inequality models. Now, we formally state the null and alternative hypotheses we consider as follows.

Test of Stochastic Dominance

$$\begin{cases} H_0^* : & \text{There exist } F^1 \text{ and } F^0 \text{ that satisfy the inequalities in Lemma 1 and} \\ & \text{such that } F^0(y) \le F^1(y) \text{ for all } y \in \mathbb{R}, \\ H_1^* : & \text{All } F^1 \text{ and } F^0 \text{ that satisfy the inequalities in Lemma 1 are such that} \\ & F^0(y) > F^1(y) \text{ for some } y \in \mathbb{R}. \end{cases}$$

When the null hypothesis H_0^* is rejected, there exist no distributions of the potential outcomes that satisfy both the model restrictions given in Lemma 1 and the stochastic dominance simultaneously. One problem is that these hypotheses cannot be tested directly since they are expressed in terms of $F^j(y)$, which are not estimable. However, they have dual representations in terms of the well-identified bounds, $L^j(y)$ and $U^j(y)$. The next lemma summarizes the dual representation result, where $\Delta(y) = L^0(y) - U^1(y)$.

Lemma 2. *The hypotheses H_0^* and H_1^* are equivalent to*

$$\begin{cases} H_0 : & \Delta(y) \le 0 \text{ for all } y \in \mathbb{R}, \\ H_1 : & \Delta(y) > 0 \text{ for some } y \in \mathbb{R}. \end{cases}$$

We have some remarks here. First, the testing problem now becomes a simple one-sided test with the object $\Delta(y)$ that can be easily estimated by a sample analog. Since the hypotheses are regarding the nonparametric object $\Delta(\cdot)$, one can apply the idea of the Kolmogorov–Smirnov test or the Cramér–von Mises test. This reasoning leads to the test statistics based on the following objects:

$$\sup_{y \in \mathbb{R}} \Delta(y) \quad \text{or} \quad \int_{\mathbb{R}} \max\{\Delta(y),\ 0\} w(y) dy, \tag{4}$$

where w is an integrable (nonnegative) weight function on $\mathbb{R}$. Both expressions in (4) are nonpositive under the null and are strictly positive under the alternative. Second, $\Delta(y)$ becomes nonpositive as $y \to \pm\infty$ so that the data contain no information for extreme values of y. Therefore, the uniformity over $\mathbb{R}$ (or the integration over $\mathbb{R}$) can be relaxed to a sufficiently large subset of $\mathbb{R}$. Also, depending on the context, one might be more interested in a specific range of y-values rather than the entire real line (e.g., income distributions below the poverty line). In such cases, one can just take supremum or integral over the restricted range of y-values of interest.

3. The Test Statistics and Asymptotic Theory

The test statistics we consider use the nonparametric sample analog of $\Delta(\cdot)$. Let $\sum_{(t_1,\cdots,t_T)}$ be the summation over all the $T!$ permutations $(t_1,\cdots,t_T)$ of $(1,2,\cdots,T)$. Using the binary indicator $1\{\cdot\}$, further let

$$1_{iT}\{y;j\} = \frac{1}{T!}\sum_{(t_1,\cdots,t_T)}\sum_{k=1}^{T} 1_i\{y;j,k\} \quad \text{and}$$
$$1_{iT}\{j\} = \frac{1}{T!}\sum_{(t_1,\cdots,t_T)} 1\{D_{it_1}=1-j,\cdots,D_{it_T}=1-j\},$$

where

$$1_i\{y;j,k\} = \begin{cases} 1\{Y_{it_1}\le y, D_{it_1}=j\} & \text{for} \quad k=1, \\ 1\{Y_{it_k}\le y, D_{it_1}=1-j,\cdots,D_{it_{k-1}}=1-j, D_{it_k}=j\} & \text{for} \quad k\ge 2. \end{cases}$$

Then, the sample analogues of $L^j(y)$ and $U^j(y)$ are given by

$$\widehat{L}^j(y) = \frac{1}{n}\sum_{i=1}^{n} 1_{iT}\{y;j\} \quad \text{and} \quad \widehat{U}^j(y) = \frac{1}{n}\sum_{i=1}^{n}\big(1_{iT}\{y;j\}+1_{iT}\{j\}\big).$$

We define $\widehat{\Delta}(y) = \widehat{L}^0(y) - \widehat{U}^0(y)$, and our statistics are given by

$$\boldsymbol{T}^{KS} = \sup_{y\in\mathbb{R}} \sqrt{n}\widehat{\Delta}(y),$$
$$\boldsymbol{T}^{CV} = \int_{\mathbb{R}} \max\{\sqrt{n}\widehat{\Delta}(y),\ 0\} w(y)dy,$$

where w is an integrable (nonnegative) weight function defined on $\mathbb{R}$. For more discussions on computation, see Appendix C.

To discuss the distributional properties of the test statistics above, we assume that the sample is independent and identically distributed (*i.i.d.*) over individuals.

Assumption 3. *The sample* $\{(Y_{it},D_{it}) : i=1,2,\cdots,n,\ t=1,2,\cdots,T\}$ *is such that* $\{(Y_i',D_i') : i=1,\cdots,n\}$ *is i.i.d., where* $Y_i=(Y_{i1},Y_{i2},\cdots,Y_{iT})'$ *and* $D_i=(D_{i1},D_{i2},\cdots,D_{iT})'$.

Let $H(y_1,y_2) = \mathrm{Cov}\big(1_{iT}\{y_1;0\}-1_{iT}\{y_1;1\}-1_{iT}\{1\},\ 1_{iT}\{y_2;0\}-1_{iT}\{y_2;1\}-1_{iT}\{1\}\big)$, and let $\mathbb{G}$ be a Gaussian process in $\ell^\infty(\mathbb{R})$ with the covariance kernel $H(\cdot,\cdot)$.[3]

[3]For any $\mathcal{Y}\subset\mathbb{R}$, $\ell^\infty(\mathcal{Y})$ is the collection of all bounded functions from $\mathcal{Y}$ to $\mathbb{R}$, and every sample path of $\mathbb{G}(\cdot)$ is an element of $\ell^\infty(\mathbb{R})$.

Theorem 1 (The Null Distributions)**.** *Suppose that Assumptions 1, 2, and 3 hold. Under H_0, there exist sequences of random variables $\{Z_n^{KS}\}$ and $\{Z_n^{CV}\}$ such that*

$$\boldsymbol{T}^{KS} \le Z_n^{KS} \xrightarrow{d} \sup_{y\in\mathbb{R}} \mathbb{G}(y),$$
$$\boldsymbol{T}^{CV} \le Z_n^{CV} \xrightarrow{d} \int_{\mathbb{R}} \max\{\mathbb{G}(y),\ 0\} w(y)dy$$

as $n \to \infty$, where the inequalities hold with probability one and they hold as equalities when $\Delta(y) = 0$ for all $y \in \mathbb{R}$.

The results in Theorem 1 do not require any regularity conditions other than the *i.i.d.* assumption and those for the identification result in Lemma 1. Since our test is a one-sided test, the inequalities in Theorem 1 are natural. Specifically, the equalities hold under the least favorable null of $\Delta(y) = 0$ for all $y \in \mathbb{R}$. Our test is not pivotal but critical values can be obtained by either simulations or the bootstrap.[4] In particular, simulating p-values can be done by adopting the p-value transformation method of Hansen (1996). The key idea is that the Gaussian process $\mathbb{G}$ can be simulated by $\hat{\mathbb{G}}_s$ defined in (5) below, where $\hat{\mathbb{G}}_s$ weakly converges in probability to $\mathbb{G}$ in $\ell^\infty(\mathbb{R})$ in the sense of Giné and Zinn (1990), for example. We illustrate the simulation procedure in more detail as follows.

Let $s \in \{1, 2, \cdots, S\}$ denote each simulation, where S is the number of replications we consider. For each s, let $\{u_{si} : i = 1, 2, \cdots, n\}$ be *i.i.d.* draws from the standard normal distribution that is independent of the data. For $y \in \mathbb{R}$, we define

$$\hat{\mathbb{G}}_s(y) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(1_{iT}\{y;0\} - 1_{iT}\{y;1\} - 1_{iT}\{1\}\right) u_{si}. \tag{5}$$

Further, we define

$$\boldsymbol{Z}_s^{KS} = \sup_{y\in\mathbb{R}} \hat{\mathbb{G}}_s(y) \quad \text{and} \quad \boldsymbol{Z}_s^{CV} = \int_{\mathbb{R}} \max\{\hat{\mathbb{G}}_s(y),\ 0\} w(y)dy.$$

Then, the simulated p-values are given by

$$\hat{p}^{KS} = \frac{1}{S}\sum_{i=1}^{S} 1\big\{\boldsymbol{Z}_s^{KS} \ge \boldsymbol{T}^{KS}\big\} \quad \text{and} \quad \hat{p}^{CV} = \frac{1}{S}\sum_{i=1}^{S} 1\big\{\boldsymbol{Z}_s^{CV} \ge \boldsymbol{T}^{CV}\big\}.$$

We now discuss the power properties of the tests. Basically all standard properties of the Kolmogorov–Smirnov test and the Cramér–von Mises test apply in our context.

[4] When w is a probability density, the integral can be computed by a Monte Carlo method, because the integral can be approximated by discretizing the support of w due to the stochastic equicontinuity of $\mathbb{G}$.

Theorem 2 (Power)**.** *Suppose that Assumptions 1, 2, and 3 hold. Let C be any constant. Then, under H_1,*

$$\mathbb{P}[\boldsymbol{T}^{KS} > C] \to 1 \quad \textit{and} \quad \mathbb{P}[\boldsymbol{T}^{CV} > C] \to 1$$

as $n \to \infty$, provided that $\sup_{y\in\mathbb{R}} \Delta(y) > 0$ and $\int_{\mathbb{R}} \max\{\Delta(y),\ 0\}w(y)dy > 0$, respectively.

Theorem 2 shows that our tests are consistent against any fixed alternatives. The consistency of the tests do not require any statistical regularity conditions either. Our tests also have non-trivial local power against $\sqrt{n}$–local alternatives. For this discussion, we consider local deviations from the least favorable null:

$$H_{1,n}: \ \Delta(y) = \delta(y)/\sqrt{n}$$

for some nonnegative function $\delta(y)$. The theorem below shows the local power properties of our tests.

Theorem 3 (Local Power)**.** *Suppose that Assumptions 1, 2, and 3 hold. Let $c^{KS}(\beta)$ and $c^{CV}(\beta)$ be critical values for size β based on the null distributions given in Theorem 1. Under $H_{1,n}$,*

$$\lim_{n\to\infty} \mathbb{P}\left[\boldsymbol{T}^{KS} > c^{KS}(\beta)\right] \geq \beta \quad \textit{and} \quad \lim_{n\to\infty} \mathbb{P}\left[\boldsymbol{T}^{CV} > c^{CV}(\beta)\right] \geq \beta,$$

where the inequalities are strict if $\sup_{y\in\mathbb{R}} \delta(y) > 0$ and $\int_{\mathbb{R}} \max\{\delta(y),\ 0\}w(y)dy > 0$, respectively.

4. Extensions

In this section we extend the proposed testing procedure in two different directions. First, we consider testing the second order stochastic dominance that gives a natural guidance for higher order cases. Second, we allow some components of A_{it} to be observed and consider the case where conditional treatment effects given observed heterogeneity can vary across different subpopulations. The null hypothesis will be the existence of a subpopulation for which the treatment has a stochastic dominance effect.

4.1. **Tests for the Second Order Stochastic Dominance.** The proposed method can be naturally extended to higher orders.To highlight the idea, we focus on the second order case in this subsection. Tests for higher orders can be dealt with similarly.

We first need the exact expressions of the hypotheses in our setup and the dual representation. Recall that $F^j(y)$ is the cdf of Y^j for $j = 0$ and 1. The standard definition says that Y^0 stochastically dominates Y^1 in the second order if and only if

$$\int_{-\infty}^{y} F^0(t)dt \leq \int_{-\infty}^{y} F^1(t)dt \quad \text{for all} \quad y \in \mathbb{R}. \tag{6}$$

Note that this hypothesis is more difficult to reject than the first order case since the former can allow for $F^0(y) > F^1(y)$ for some y and is implied by the latter. We want to test the hypothesis in (6) while incorporating the partial identification given in Lemma 1. The same reasoning in the first order case leads us into the following hypotheses representation.

Test of Second Order Stochastic Dominance

$$\begin{cases} H_{0S}^* : \text{ There exist } F^1(\cdot) \text{ and } F^0(\cdot) \text{ that satisfy the inequalities in Lemma 1 and} \\ \qquad \text{such that } \int_{-\infty}^{y} F^0(t)dt \leq \int_{-\infty}^{y} F^1(t)dt \text{ for all } y \in \mathbb{R}, \\ H_{1S}^* : \text{ All } F^1(\cdot) \text{ and } F^0(\cdot) \text{ that satisfy the inequalities in Lemma 1 are such that} \\ \qquad \int_{-\infty}^{y} F^0(t)dt > \int_{-\infty}^{y} F^1(t)dt \text{ for some } y \in \mathbb{R}. \end{cases}$$

To introduce the dual representation, we first assume that Y_{it}^0 and Y_{it}^1 have compact support, of which the union will be denoted by $\mathcal{Y} \subset \mathbb{R}$. We do this because $\Delta(y)$ could become a strictly negative constant as $y \to \pm\infty$ so that integrating $\Delta(\cdot)$ from $-\infty$ is not generally well-defined. However, this assumption is not a serious limitation, because $F^0(y)$ and $F^1(y)$ are trivially point–identified as 0 or 1 whenever $y \to \pm\infty$. A similar assumption can be also found in Barrett and Donald (2003).

Assumption 4. *Y_{it}^0 and Y_{it}^1 have compact support so that their union, denoted by $\mathcal{Y} \subset \mathbb{R}$, is also compact.*

Under assumption 4, the proposed hypotheses can be represented by

$$\begin{cases} H_{0S} : \int_{y_{\min}}^{y} \Delta(t)dt \leq 0 \quad \text{for all} \quad y \in \mathcal{Y}, \\ H_{1S} : \int_{y_{\min}}^{y} \Delta(t)dt > 0 \quad \text{for some} \quad y \in \mathcal{Y}, \end{cases}$$

where $y_{\min} = \inf \mathcal{Y}$. The duality between H_{0S}^*, H_{1S}^* and H_{0S}, H_{1S} follows similarly to Lemma 2 since integration is a monotone operator. Now we can define the test statistics similarly as the first order case:

$$\boldsymbol{T}_S^{KS} = \sup_{y \in \mathcal{Y}} \int_{y_{\min}}^{y} \sqrt{n}\widehat{\Delta}(t)dt, \quad \boldsymbol{T}_S^{CV} = \int_{\mathbb{R}} \max\{\int_{y_{\min}}^{y} \sqrt{n}\widehat{\Delta}(t)dt,\ 0\}w(y)dy,$$

where w is an integrable (nonnegative) weight function on $\mathcal{Y} \subset \mathbb{R}$. We summarize its null distribution in the next theorem.

Theorem 4. *Suppose that Assumptions 1, 2, and 3 hold. Under H_{0S}, there exist sequences of random variables $\{Z_{S,n}^{KS}\}$ and $\{Z_{S,n}^{CV}\}$ such that*

$$\boldsymbol{T}_S^{KS} \leq Z_{S,n}^{KS} \xrightarrow{d} \sup_{y\in\mathcal{Y}} \int_{y_{min}}^{y} \mathbb{G}(t)dt,$$
$$\boldsymbol{T}_S^{CV} \leq Z_{S,n}^{CV} \xrightarrow{d} \int_{\mathcal{Y}} \max\{\int_{y_{min}}^{y} \mathbb{G}(t)dt,\ 0\}w(y)dy$$

as $n \to \infty$, where the inequalities hold with probability one and they hold as equalities when $\Delta(y) = 0$ for all $y \in \mathcal{Y}$.

For the power and local power properties, Theorems 2 and 3 can be similarly extended.

4.2. **Observed Heterogeneity.** Suppose now that the heterogeneity A_{it}, which was needed to be controlled for to achieve randomization, consists of two parts: a vector of observable covariates X_{it} and a vector of unobserved components α_{it}. Treatment effects could be heterogeneous over different subgroups characterized by X_{it}. For example, some workers may have positive wage premium of the union membership while the others have negative premium. If the absolute magnitude of these two effects are similar, a simple test may mislead a researcher to conclude that there is no treatment effect even though there exist substantial *conditional* treatment effects.

Crump, Hotz, Imbens, and Mitnik (2008) and Lee and Whang (2009) proposed testing procedures for these heterogeneous (or conditional) treatment effects under the standard unconfoundedness assumption (or selection on *observables*). Since we do not make the unconfoundedness assumption, their approaches are not directly applicable. However, our specification testing approach based on a set identification can be modified to fit their framework. We pay special attention to Lee and Whang (2009) in this subsection, because the limit distribution of the test statistic with an extra integration over X_{it} turns out to be a normal distribution.

Below we focus on the first order dominance case and we show how to formulate the problem without discussing too much technicality.[5] Before proceeding, we modify Assumption 3 to include the observed covariates X_{it}.

Assumption 5. *The sample $\{(Y_{it}, D_{it}, X_{it}')' : \ i = 1, 2, \cdots, n, \ t = 1, 2, \cdots, T\}$ is such that $\{(Y_i', D_i', X_i')'\}$ is i.i.d., where $Y_i = (Y_{i1}, Y_{i2}, \cdots, Y_{iT})'$, $D_i = (D_{i1}, D_{i2}, \cdots, D_{iT})'$, and $X_i = (X_{i1}', X_{i2}', \cdots, X_{iT}')'$. We assume that X_i is continuous and has a probability density function f.*

[5]A formal result is stated below, but technical assumptions similar to Lee and Whang (2009) are given in Appendix B.7.

Let $A_{it} = (X'_{it}, \alpha'_{it})'$ and $\alpha_i = (\alpha'_{i1}, \cdots, \alpha'_{iT})'$, where X_{it} and α_{it} are defined as above. We consider the conditional distributions of the potential outcomes as before:

$$F^j(y|x) = \int \mathbb{P}\left[Y_{it}^j \leq y \big| X_i = x, \alpha_i = a\right] d\mu_{\alpha_i|X_i}(a|x) \quad \text{for} \quad j = 0, 1, \tag{7}$$

where $x \in \mathcal{X}$ with $\mathcal{X}$ being the support of X_i.

We now ask if there is any evidence that shows that the treatment is effective at least for some subpopulation $\mathcal{X}_s$ in $\mathcal{X}$. The lack of point-identification of $F^j(y|x)$ makes this problem nonstandard and different from Lee and Whang (2009). However, the conditional version of Lemma 1 is available under Assumptions 1 and 2, from which we take the specification testing approach as in Sections 2 and 3.

Let $p_k^j(y|x)$ be defined as $p_k^j(y)$ but with conditioning on $X_i = x$, and let

$$L^j(y|x) = \sum_{k=1}^{T} p_k^j(y|x),$$
$$U^j(y|x) = L^j(y|x) + \mathbb{P}[D_{it_1} = 1 - j, D_{it_2} = 1 - j, \cdots, D_{it_T} = 1 - j | X_i = x].$$

The same calculation of Lemma 1 with conditioning on $X_i = x$ gives that for $j = 0$ and 1,

$$L^j(y|x) \leq F^j(y|x) \leq U^j(y|x) \text{ for all } (y, x) \in \mathcal{Y} \times \mathcal{X}_s, \tag{8}$$

where $\mathcal{Y}$ is the support of y and $\mathcal{X}_s \subset \mathcal{X}$. The proof for these bounds is omitted since it is identical to that of Lemma 1. As in the previous sections, we treat the inequalities in (8) as model restrictions that should hold always. We then test whether the existence of subpopulation, for which the first order stochastic dominance holds, is compatible with the model restrictions. More precisely, we consider the following hypotheses of the conditional treatment effect:

$$\begin{cases} H_{0C}^*: \text{ There exist } F^1(\cdot|\cdot) \text{ and } F^0(\cdot|\cdot) \text{ that satisfy inequalities in (8) and} \\ \qquad \text{such that } F^0(y|x) \leq F^1(y|x) \text{ for all } y \in \mathcal{Y} \text{ and all } x \in \mathcal{X}_s, \\ H_{1C}^*: \text{ All } F^1(\cdot|\cdot) \text{ and } F^0(\cdot|\cdot) \text{ that satisfy the inequalities in (8) are such that} \\ \qquad F^0(y|x) > F^1(y|x) \text{ for some } y \in \mathcal{Y} \text{ and all } x \in \mathcal{X}_s, \end{cases}$$

which are equivalent to

$$\begin{cases} H_{0C}: \ \Delta(y, x) \leq 0 \quad \text{for all} \quad (y, x) \in \mathcal{Y} \times \mathcal{X}_s, \\ H_{1C}: \ \Delta(y, x) > 0 \quad \text{for some} \quad (y, x) \in \mathcal{Y} \times \mathcal{X}_s, \end{cases} \tag{9}$$

where $\Delta(y, x) = L^0(y|x) - U^1(y|x)$.

The (dual) hypotheses in (9) are the conditional versions of those in Lemma 2. When the null hypothesis is not rejected, it is interpreted as there is no evidence against the first order stochastic dominance for any subpopulation $\mathcal{X}_s$ in $\mathcal{X}$. Note that the hypotheses in (9) are now comparable with the ones analyzed in Lee and Whang (2009).

We now define our test statistic. In principle, one can think of both sup-type and integration-type tests, but we only focus on the integration approach here.[6] Additional notation is required. Let $K(\cdot)$ be a kernel function with compact support, and let $K_h(s) = K(s/h_n)/h_n^{d_T}$, where $d_T = \sum_{t=1}^{T} \dim(X_{it})$ and h_n is a bandwidth choice. Let

$$\widehat{L}^j(y|x) = \frac{1}{\hat{f}(x)} \frac{1}{n} \sum_{i=1}^{n} 1_{iT}\{y;j\} K_h(x - X_i),$$

$$\widehat{U}^j(y|x) = \frac{1}{\hat{f}(x)} \frac{1}{n} \sum_{i=1}^{n} (1_{iT}\{y;j\} + 1_{iT}\{j\}) K_h(x - X_i),$$

where $\widehat{f}(x) = n^{-1} \sum_{i=1}^{n} K_h(x - X_i)$. Further let $\widehat{\Delta}(y,x) = \widehat{L}^0(y|x) - \widehat{U}^1(y|x)$. Then, we define the statistic $\boldsymbol{T}^C$ for the conditional treatment effect by

$$\boldsymbol{T}^C = \int \int \max\{\sqrt{n}\widehat{\Delta}(y,x),\ 0\} \tilde{w}(y,x) dy dx,$$

where $\tilde{w}$ is a weight function whose support is a compact set $\mathcal{Y} \times \mathcal{X}_s$. The statistic $\boldsymbol{T}^C$ is similar to $\boldsymbol{T}^{CV}$ except that we have an extra integral over X_i. In fact, this extra integral is what makes the null distribution of $\boldsymbol{T}^C$ completely different from that of $\boldsymbol{T}^{CV}$. The Poissonization technique shows that it is a normal distribution.

To characterize the asymptotic properties of $\boldsymbol{T}^C$, we need the following functions:

$$\begin{aligned} q_i(y) &= 1_{iT}\{y;0\} - 1_{iT}\{y;1\} - 1_{iT}\{1\} \\ F(\rho) &= \mathrm{Cov}\left(\max\{\sqrt{1-\rho}\mathbb{Z}_1 + \rho\mathbb{Z}_2,\ \max\{\mathbb{Z}_2,\ 0\}\right), \\ \rho_v(y,x) &= \frac{\mathrm{Var}\left(q_i(y)|X_i = x\right)}{f(x)}, \\ \rho(y,\tilde{y},x,t) &= \frac{K_*(t)}{K_*(0)} \frac{\mathrm{Cov}(q_i(y), q_i(\tilde{y})|X_i = x)}{f(x)\sqrt{\rho_v(y,x)\rho_v(\tilde{y},x)}}, \end{aligned}$$

where $\mathbb{Z}_1$ and $\mathbb{Z}_2$ are independent standard normal random variables, and $K_*(t) = \int K(u)K(u+t)du$. Theorem 5 summarizes the properties of $\boldsymbol{T}^C$ under the null, alternative, and local alternatives, which directly follow from Lee and Whang (2009).

[6] To the best of our knowledge, the asymptotic distribution of the sup–based test for (9) is an open question. See also Lee and Whang (2009).

Theorem 5. *Suppose that Assumptions 1, 2, 5, and 6 hold.*[7] *Under the least favorable case of* H_{0C}*, i.e.,* $\Delta(y|x) = 0$ *for all* $(y,x) \in \mathcal{Y} \times \mathcal{X}_s$*,*

$$\frac{\widehat{\boldsymbol{T}}^{C} - \eta_n}{\sigma_0} \xrightarrow{d} N(0,1)$$

as $n \to \infty$*, where*

$$\eta_n = h^{-d_T/2} \int_{\mathbb{R}^{d_T}} \int_{\mathbb{R}} \sqrt{\rho_v(y,x)} \widetilde{w}(y,x)\, dy dx \cdot \mathbb{E}\left[\max\{\mathbb{Z}_1, 0\}\right],$$

$$\sigma_0^2 = \int_{\{\|t\| \le 1\}} \int_{\mathbb{R}^{d_T}} \int_{\mathbb{R}} \int_{\mathbb{R}} F\left[\rho(y,\tilde{y},x,t)\right] \sqrt{\rho_v(y,x)\, \rho_v(\tilde{y},x)} \tilde{w}(y,x) \tilde{w}(\tilde{y},x) dy d\tilde{y} dx dt.$$

To show the consistency, let C *be any constant. Then, under the alternative hypothesis* H_{1C}*,*

$$\mathbb{P}\left(\frac{\widehat{\boldsymbol{T}}^{C} - \eta_n}{\sigma_0} > C \right) \to 1$$

as $n \to \infty$*. Finally, let* $\delta(\cdot,\cdot)$ *be a real non-negative function, and define local alternatives by* $H_{1C,n} : \Delta(y,x) = n^{-1/2}\delta(y,x)$*. Then, under the local alternatives* $H_{1C,n}$*,*

$$\frac{\widehat{\boldsymbol{T}}^{C} - \eta_n}{\sigma_0} \xrightarrow{d} N(0,1) + c$$

as $n \to \infty$ *for a positive constant* c*.*

Theorem 5 does not deal with the fact that the bias and standard error should be estimated in practice. Since this issue is discussed in detail in Lee and Whang (2009), we do not repeat it here.

5. Empirical Illustration: The Effect of Smoking on Birth Outcomes

In this section, we study the causal effect of smoking during pregnancy on infant's birthweight. This question has been investigated in health economics (and in econometrics), and many studies provide evidence that smoking indeed decreases infant's birthweight; e.g., Permutt and Hebel (1989), Evans and Ringel (1999), Abrevaya (2006), Abrevaya and Dahl (2008), and Hoderlein and Sasaki (2011). However, there are still remaining concerns such as the presence of unobserved heterogeneity and the validity of instrumental variables.

Mother's smoking behavior is correlated with other life-style factors, which accounts for the importance of unobserved heterogeneity. To resolve this endogeneity problem, most studies in the existing literature including those cited above use either an instrumental variable (IV) estimator or a fixed-effect estimator. For the IV approach, however, it is

[7] Assumption 6 is a technical one that follows Lee and Whang (2009). The assumption and comments on it are provided in Appendix B.7.

not always obvious to obtain good instrumental variables; large confidence region of the IV estimator makes it difficult to statistically distinguish the result from that of the OLS estimator. Alternatively, when panel data are available, fixed-effect estimation might be a natural approach. However, when a mother quits smoking, her (unobserved) health-related habits might also change together. Thus, the time-invariance of fixed-effects seems to be a strong assumption in this context.

We relax such restrictions in our empirical study by using the method developed in the previous sections. Specifically, in a panel data setup, we allow for *time-varying* unobserved heterogeneity that can be *nonseparable* to the other variables. With this flexible model, we test the presence of first-order stochastic dominance relationship between potential birthweight outcomes.

Our analysis is based on the data set constructed by Abrevaya (2006) from the U.S. federal natality data in 1990–1998. Since the original data do not provide unique identifiers, he applied several matching algorithms to find out each pair of a mother and children. We select the "matched panel #3" as it is constructed in the most conservative way. The same data set is also used by Arellano and Bonhomme (2010) in the random coefficients panel model.

For notation, Y_{it} is the birthweight of the t^{th} baby of mother i; $D_{it} = 1$ if the mother smoked during the pregnancy and 0, otherwise; Y_{it}^1 and Y_{it}^0 are potential birthweights depending on the mother's smoking status.

Table 1 summarizes the proportion of smoking behavior during pregnancy when each mother has two or three children, respectively. In either case, more than 82% of mothers had never smoked. We focus on the subpopulation of those who had ever smoked during pregnancy for the following reasons. First, the sample size of those who had never smoked is too large to obtain any meaningful bounds. Second, those who ever smoked make a more relevant population under the presumption that smokers may quit in the future but non-smokers are unlikely to start smoking during pregnancy. As a result, the distribution function $F^1(\cdot)$ is point-identified while the other distribution $F^0(\cdot)$ is only partially identified.

Figures 1–3 report the estimated distributions of $F^0(\cdot)$ and $F^1(\cdot)$ in *Ever-Smoker* sample of three births and in the various subsamples. In each graph, the solid line stands for the estimates of $F^1(\cdot)$, and the shaded area for $F^0(\cdot)$. Looking at the upper panel of Figure 1, which is from *3-Birth*, we can find that $\widehat{F}^1(\cdot)$ is located above the lower bound of $\widehat{F}^0(\cdot)$ over all birthweights. This implies that there might exist F^0 stochastically dominating F^1 in the first order. However, $\widehat{F}^1(\cdot)$ clearly crosses the upper bound of $\widehat{F}^0(\cdot)$ around the birthweight of $4,000$ grams, which implies that we might not find $F^0(\cdot)$ stochastically dominated by $F^1(\cdot)$ in the identified region.

Table 1. Proportions of Smoking Behavior

Two Births ($T = 2$)				
Full sample ($n = 141,929$)			Ever smoker ($n = 25,390$)	
Never smoker	Switcher	Always smoker	Switcher	Always smoker
82.1%	9.8%	8.1%	54.5%	45.5%

Three Births ($T = 3$)				
Full sample ($n = 12,360$)			Ever smoker ($n = 2,137$)	
Never smoker	Switcher	Always smoker	Switcher	Always smoker
82.7%	5.6%	11.7%	67.6%	32.4%

This graphical intuition is formally confirmed by the statistical test results in Table 2. We first test the null hypothesis that the distribution of birthweight from a nonsmoking mother first-order stochastically dominates that from a smoking mother (F^0 FSD F^1). We next test the null of the other direction (F^1 FSD F^0). At the row of *3-Birth*, both the Kolmogorov–Smirnov ($\boldsymbol{T}^{KS}$) test and the Cramér–von Mises ($\boldsymbol{T}^{CV}$) test cannot reject the first null but strongly reject the flipped one. We also report the test results with the sample of two births at the row of *2-Birth* in the same table. Only the $\boldsymbol{T}^{CV}$ test yields similar results to those of *3-Birth*, and the $\boldsymbol{T}^{KS}$ test cannot reject the flipped null. This shows that an additional time period helps the test perform better by improving identification power.

We next turn our attention to low and high quantiles of the distributions. Specifically, we consider the cases where the minimum birthweight among three children from the same mother is below 2,500 grams ($BW_{min} < 2,500g$) or the maximum is above 4,080 grams ($BW_{max} > 4,080$). The weights of 2,500 grams and 4,080 grams correspond to the 10^{th} and 90^{th} percentiles of the birthweight distribution in the U.S., which are usually used to define 'underweighted' and 'overweighted' infants. The middle and lower panels in Figure 1 show that the estimated distributions have the same patterns as above, and the statistical tests in Table 2 again confirm it. Thus, we may conclude that the same relationship (F^0 FSD F^1) holds conditional on low and high quantiles of the birthweight distributions.

We also consider various subpopulations. Based on mother's characteristics in the data, we select the following subgroups: *No Marriage*, *No High School*, *No Alcohol*, *All Girls*, and *All Boys*, where all group names are self-explanatory. (Detailed definitions are given under Table 2.) Figures 2–3 and Table 2 summarize the results for these subpopulations. Overall, the results are consistent with those of the ever-smoker population except *All Boys*. In the subpopulations of *No Marriage* and *No High School*, the estimated distributions $\widehat{F}^1(\cdot)$ are located slightly above the upper bounds of $\widehat{F}^0(\cdot)$ over short intervals, which yield relatively high p-values in the $\boldsymbol{T}^{CV}$ test (10.4% and 13.4%, respectively). In the subpopulation of *All Boys*, the estimate $\widehat{F}^1(\cdot)$ crosses both with the lower bound (around $1,500$ grams) and with the upper bound (around $4,000$ grams) of $\widehat{F}^0(\cdot)$. The tests in Table 2 also reject the null hypotheses in both directions, and do not confirm any conclusive stochastic relationship.

Finally, we repeat all statistical tests listed above with the mean-adjusted samples. There is a tendency that, among siblings, infants born in the later order are heavier than those born earlier, which suggests that the time homogeneity assumption may be too strong. In fact, the second and third infants in the sample are heavier by about 74 and 80 grams, respectively, on average than the first one. Table 3 summarizes the test results with the mean-adjusted sample by these mean shifts. Overall, the results are similar to those from the original sample. In the subpopulations of *No Marriage* and *No High School*, it turns out that the p-values of the $\boldsymbol{T}^{CV}$ test even decreases.

To sum up, this empirical study provides more rigorous scientific evidence that smoking during pregnancy negatively affects birthweights. Allowing for nonseparable and time-varying unobserved heterogeneity, it confirms that the distribution of the birthweight from a smoking mother might be stochastically dominated by that from a nonsmoking mother.

6. Conclusions

In this paper we proposed testing methods for stochastic dominance between the distributions of potential outcomes. Not relying on the unconfoundedness assumption, we allowed for the presence of unobserved heterogeneity. We considered the availability of panel data, but, unlike the standard panel literature, we do not assume the time-invariance of unobserved heterogeneity. Despite this level of generality nontrivial bounds of the distributions of potential outcomes are still available, and we showed how to use them to test stochastic dominance among the potential outcome distributions. Using the proposed methods, we investigated the causal effects of smoking during pregnancy on infant's birthweight, and reinforced the empirical evidence showing that smoking negatively affects birth outcomes.

We conclude this paper by suggesting areas of future research. First, the proposed methods can be immediately applied to many empirical studies on treatment effects with a lot of flexibility. Second, an idea similar to our partial identification approach can be applied to other contexts such as testing for monotonicity or structural changes when the model is partially identified. Finally it is also worth deriving an asymptotically admissible test within the set identification setup.

Appendix A. Tables and Figures

Table 2. p-values and Sample Sizes of the Original Sample

(*Null Hypothesis*)	F^0 FSD F^1		F^1 FSD F^0		no. obs.	
	T^{KS}	T^{CV}	T^{KS}	T^{CV}	Switcher	Always Smoker
2-Birth	1.000	0.521	0.320	0.046	13,846	11,544
3-Birth	1.000	0.538	0.000	0.000	1,445	692
Subpopulations of *3-Birth*						
$BW_{min} < 2,500g$	1.000	0.519	0.000	0.000	251	167
$BW_{max} > 4,080g$	1.000	0.486	0.000	0.000	294	65
No Marriage	0.802	0.486	0.000	0.103	512	296
No High School	1.000	0.520	0.000	0.134	694	399
No Alcohol	1.000	0.499	0.000	0.000	1,265	570
All Girls	1.000	0.505	0.000	0.000	173	109
All Boys	0.000	0.111	0.000	0.000	218	92

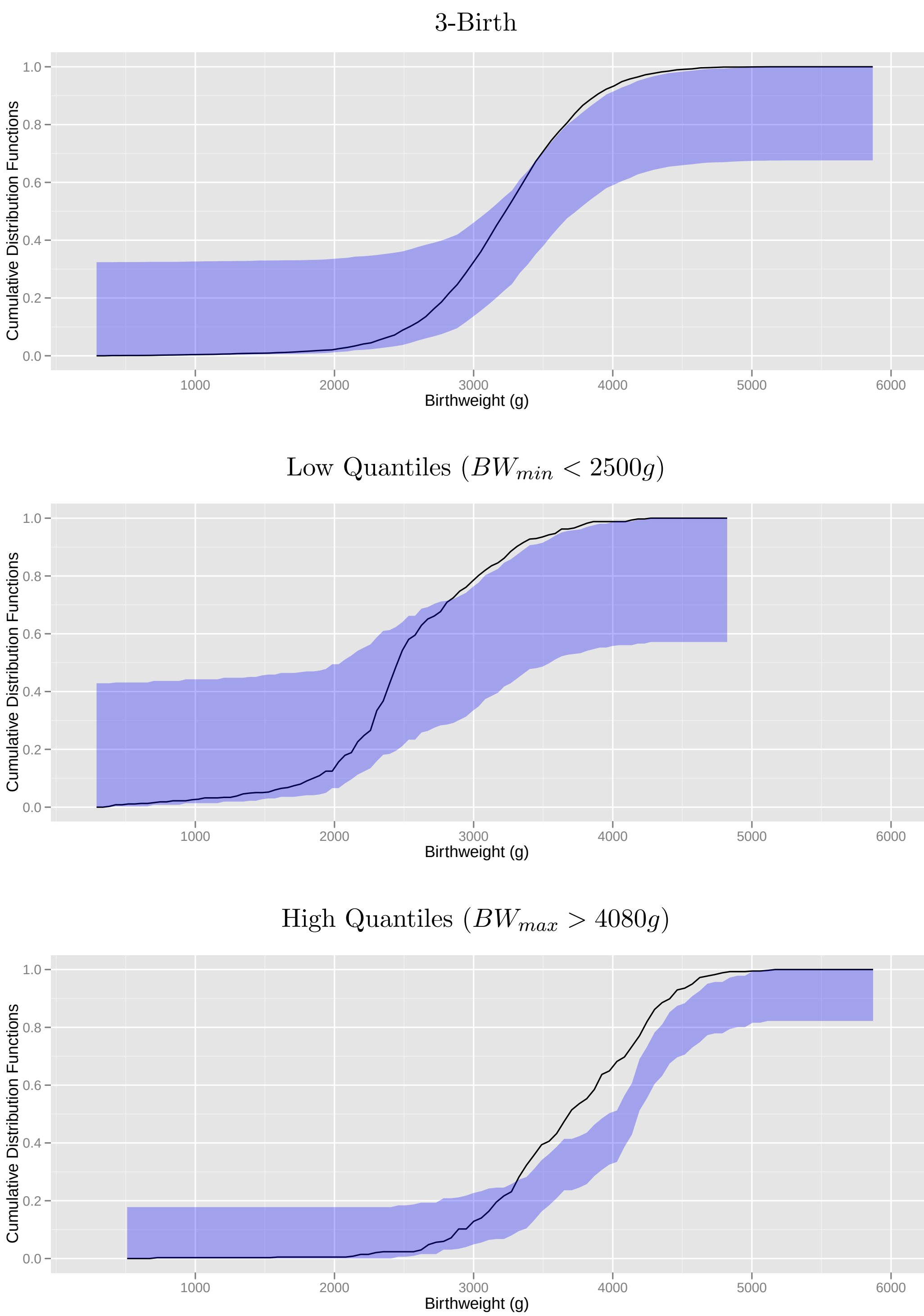
Note: FSD stands for 'First-order Stochastically Dominates.' T^{KS} is the Kolmogorov-Smirnov type test and T^{CV} is the Cramér-von Mises type test. *2-Birth* and *3-Birth* stand for 2-births and 3-births in *Ever-Smoker* sample, respectively. The remaining subpopulations are from *3-Birth*. BW_{min} stands for the minimum birthweight among the three children from the same mother, and BW_{max} is defined similarly. The definitions of each subpopulation are as follows. *No Marriage*: mothers who never get married; *No High School*: mothers who do not graduate high schools; *No Alcohol*: mothers who never drink during three pregnancies; *All Girls*: mothers whose three children are all female; *All Boys*: mothers whose three children are all male.

Table 3. p-values and Sample Sizes of the Mean-adjusted Sample

(*Null Hypothesis*)	F^0 FSD F^1		F^1 FSD F^0		no. obs.	
	T^{KS}	T^{CV}	T^{KS}	T^{CV}	Switcher	Always Smoker
2-Birth	1.000	0.516	0.336	0.055	13,846	11,544
3-Birth	0.976	0.538	0.000	0.000	1,445	692
Subpopulations of *3-Birth*						
$BW_{min} < 2,500g$	1.000	0.520	0.000	0.000	251	167
$BW_{max} > 4,080g$	1.000	0.493	0.000	0.000	294	65
No Marriage	0.805	0.486	0.000	0.066	512	296
No High School	1.000	0.521	0.000	0.118	694	399
No Alcohol	0.916	0.499	0.000	0.000	1,265	570
All Girls	1.000	0.510	0.000	0.000	173	109
All Boys	0.000	0.071	0.000	0.000	218	92

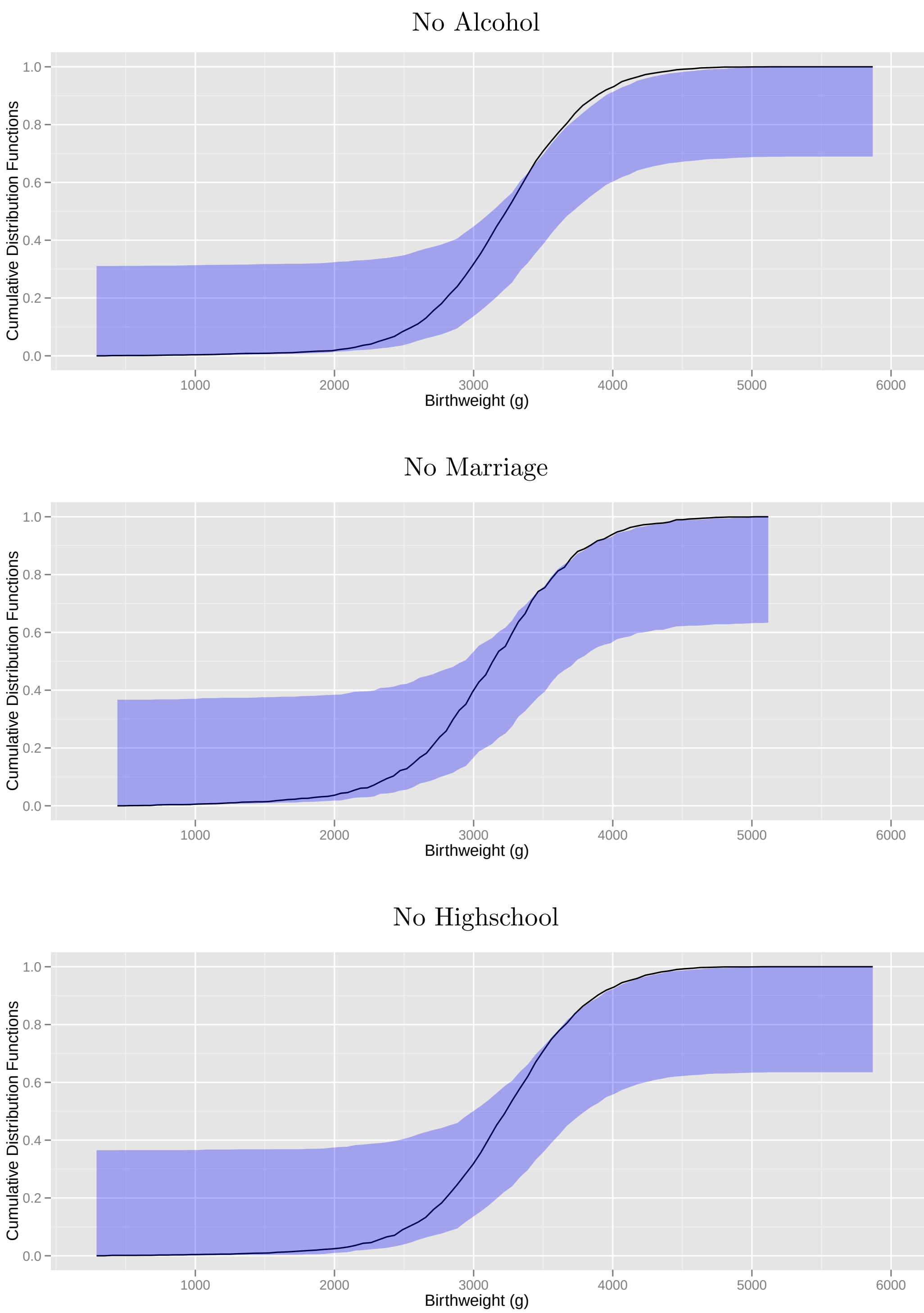
Note: The original sample is mean-adjusted by the amounts of mean shifts in $T = 2$ and 3. See the note below Table 2 for other notation.

FIGURE 1. Distributions of Birthweight



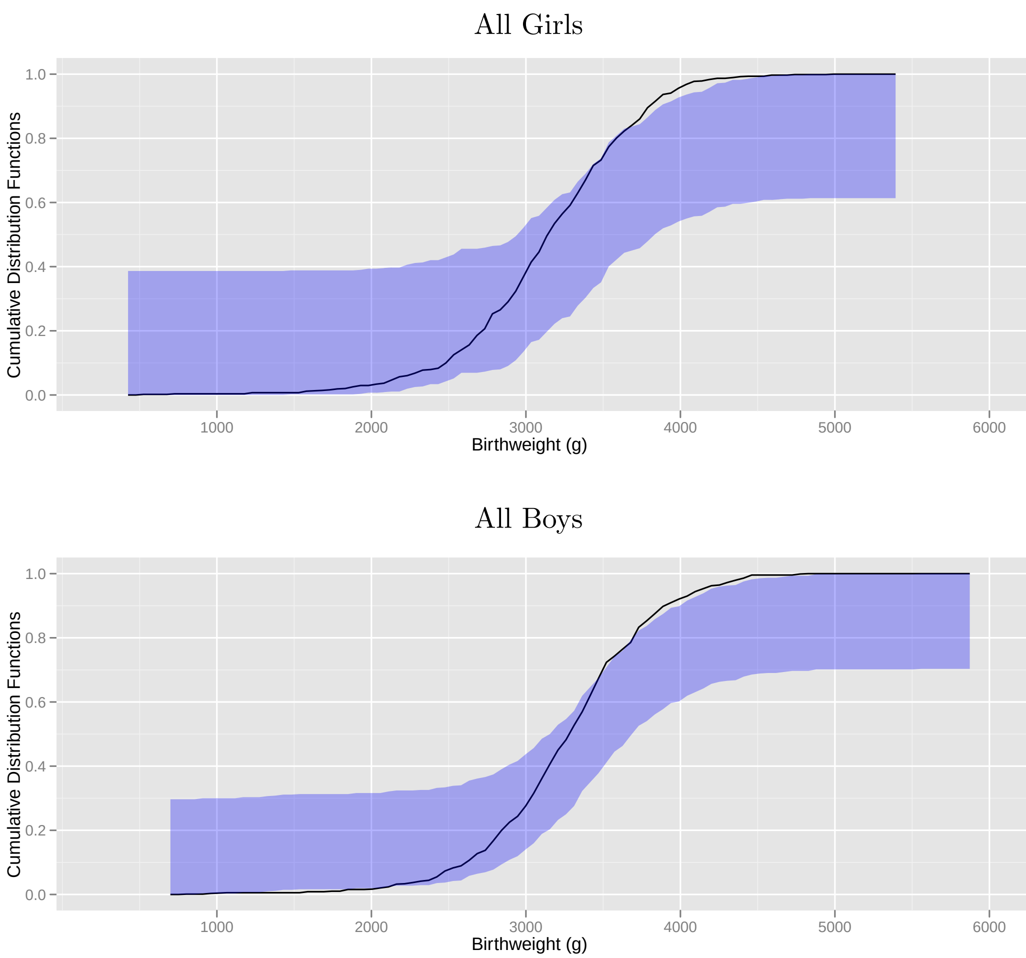
Note: The solid lines stand for the estimates of $F^1(\cdot)$, and the shaded areas for $F^0(\cdot)$. BW_{min} is the minimum birthweight among three children from the same mother, and BW_{max} is defined similarly.

FIGURE 2. Distributions of Birthweight in Subpopulations



Note: The solid lines stand for the estimates of $F^1(\cdot)$, and the shaded areas for $F^0(\cdot)$.

FIGURE 3. Distributions of Birthweight in Subpopulations



Note: The solid lines stand for the estimates of $F^1(\cdot)$, and the shaded areas for $F^0(\cdot)$.

Appendix B. Proofs

Throughout the appendix we use the abbreviations RHS and LHS for *right–hand side* and *left–hand side*, respectively. Further, RHS1 means the first term on the right–hand, and similarly for RHS2, RHS3, etc..

B.1. **Proof of Lemma 1.** We will focus on $j = 1$ since the other case is symmetric. Fix a particular permutation $(t_1, t_2, \cdots, t_T)$. Recall from (2) that

$$F^1(y) = \mathbb{P}\left[D_{it_1} = 0\right] \int \mathbb{P}\left[Y^1_{it_1} \leq y \middle| A_i = a\right] d\mu_{A_i|D_{it_1}}(a|0) + \mathbb{P}\left[Y_{it_1} \leq y,\ D_{it_1} = 1\right], \quad (10)$$

where the RHS2 is $p^1_1(y)$. The RHS1 of (10) is equal to the sum of the two terms:

$$\mathbb{P}\left[D_{it_1} = 0, D_{it_2} = 0\right] \int \mathbb{P}\left[Y^1_{it_1} \leq y \middle| A_i = a\right] d\mu_{A_i|D_{it_1},D_{it_2}}(a|0,0), \quad (11)$$

$$\mathbb{P}\left[D_{it_1} = 0, D_{it_2} = 1\right] \int \mathbb{P}\left[Y^1_{it_1} \leq y \middle| A_i = a\right] d\mu_{A_i|D_{it_1},D_{it_2}}\left(a|0,1\right), \quad (12)$$

where term (12) is equal to $p^1_2(y)$ by Assumptions 1 and 2.

Therefore, the RHS of (10) is now the sum of (11) and $\sum_{k=1}^{2} p^1_k(y)$. Apply the same calculation to (11) and continue until we reach t_T. We then obtain

$$\begin{aligned} F^1(y) &= \sum_{k=1}^{T} p^1_k(y) \\ &\quad + \mathbb{P}\left[D_{it_1} = 0, \cdots, D_{it_T} = 0\right] \int \mathbb{P}\left[Y^1_{it_1} \leq y \middle| A_i = a\right] d\mu_{A_i|D_{it_1},\cdots,D_{it_T}}(a|0,\cdots,0), \end{aligned}$$

from which the statement follows. □

B.2. **Proof of Lemma 2.** We first need some notation. For each $y \in \mathbb{R}$, define the bounds of the true underlying distributions:

$$\mathcal{F}(y) = \left\{(a, b) \in \mathbb{R}^2 :\ L^0(y) \leq a \leq U^0(y) \quad \text{and} \quad L^1(y) \leq b \leq U^1(y)\right\},$$

which is not empty for any $y \in \mathbb{R}$ by construction. Also, note that $0 \leq L^j(y) \leq U^j(y) \leq 1$ for all $y \in \mathbb{R}$ and $j = 0, 1$, where $U^j(y) - L^j(y) = \mathbb{P}[D_{it_1} = 1 - j, D_{it_2} = 1 - j, \cdots, D_{it_T} = 1 - j] = \mathcal{P}^j$. Further let

$$\mathcal{G} = \left\{(a, b) \in \mathbb{R}^2 :\ a \leq b,\ 0 \leq a \leq 1,\ 0 \leq b \leq 1\right\}.$$

• **Equivalence of the Null Hypotheses:**

(i) Sufficiency: Let H^*_0 be true. Then, there exist distribution functions F^0 and F^1 such that for any $y \in \mathbb{R}$, $\left(F^0(y), F^1(y)\right) \in \mathcal{F}(y) \cap \mathcal{G}$. Fix such F^0 and F^1 and suppose that H_0 is not true. Then, there exists $\tilde{y} \in \mathbb{R}$ such that $\Delta(\tilde{y}) = L^0(\tilde{y}) - U^1(\tilde{y}) > 0$, which implies that $F^1(\tilde{y}) < F^0(\tilde{y})$. This contradicts to $\big(F^0(\tilde{y}), F^1(\tilde{y})\big) \in \mathcal{G}$.

(ii) Necessity: Let H_0 be true. Then, $\mathcal{F}(y) \cap \mathcal{G} \neq \emptyset$ for all $y \in \mathbb{R}$. We will construct distribution functions F^0 and F^1 such that $\big(F^0(y), F^1(y)\big) \in \mathcal{F}(y) \cap \mathcal{G}$ for all $y \in \mathbb{R}$. Fix a sufficiently small $\epsilon > 0$, and we define F^0 and F^1 as follows: (a) $0 \leq L^1(y) < \epsilon$, let $F^1(y) = \max\{L^0(y), L^1(y)\}$; (b) $1 - \epsilon < U^0(y) \leq 1$, let $F^0(y) = \min\{U^0(y), U^1(y)\}$; (c) otherwise, let $F^0(y) = L^0(y)$ and $F^1(y) = U^1(y)$. Then, $\big(F^0(y), F^1(y)\big) \in \mathcal{F}(y) \cap \mathcal{G}$ for all $y \in \mathbb{R}$ by construction. Since F_0, F^1 are CADLAG functions by construction, they are distribution functions.

• **Equivalence of the Alternative hypotheses:**

(i) Sufficiency: Let H_1^* be true. Fix $M > 0$, and choose two distribution functions F^0 and F^1 such that (a) when $-M \leq y < M$, $\big(F^0(y), F^1(y)\big) = \big(L^0(y), U^1(y)\big)$; and (b) otherwise, $F^0(y) = F^1(y) \to 0$ as $y \to \infty$ and $F^0(y) = F^1(y) \to 1$ as $y \to \infty$. We then have $\big(F^0(y), F^1(y)\big) \in \mathcal{F}(y)$ for all $y \in \mathbb{R}$ by construction, provided that M is sufficiently large. Therefore, there must be some $\tilde{y} \in \mathbb{R}$ such that $F^0(\tilde{y}) > F^1(\tilde{y})$, which implies that $\tilde{y} \in [-M, M]$. So, $\tilde{y}$ satisfies $F^0(\tilde{y}) = L^0(\tilde{y}) > U^1(\tilde{y}) = F^1(\tilde{y})$, which implies that $\Delta(\tilde{y}) > 0$.

(ii) Necessity: Let H_1 be true. Then, there exists $\tilde{y} \in \mathbb{R}$ such that $\mathcal{F}(\tilde{y}) \subset \mathcal{G}^c$, where $\mathcal{G}^c = \{(a, b) \in \mathbb{R}^2 :\ a > b,\ 0 \leq a \leq 1,\ 0 \leq b \leq 1\}$. Consider any distribution functions F^0 and F^1 that satisfy $\big(F^0(y), F^1(y)\big) \in \mathcal{F}(y)$ for all $y \in \mathbb{R}$. It then follows that $\big(F^0(\tilde{y}), F^1(\tilde{y})\big) \in \mathcal{F}(\tilde{y}) \subset \mathcal{G}^c$. Therefore, $F^0(\tilde{y}) > F^1(\tilde{y})$ and H_1^* follows. □

Remark: Figure 4 illustrates the hypotheses in Lemma 2.

B.3. **Proof of Theorem 1.** Note that under H_0, for all $y \in \mathbb{R}$,

$$\sqrt{n}\widehat{\Delta}(y) \leq \sqrt{n}\left(\widehat{\Delta}(y) - \Delta(y)\right),$$

$$\max\{\sqrt{n}\widehat{\Delta}(y),\ 0\} \leq \max\{\sqrt{n}\left(\widehat{\Delta}(y) - \Delta(y)\right),\ 0\},$$

where the inequalities become equalities when $\Delta(y) = 0$ for all $y \in \mathbb{R}$. Therefore, by the continuous mapping theorem, it suffices to show that $\sqrt{n}(\widehat{\Delta} - \Delta)$ weakly converges to $\mathbb{G}$ in $\ell^\infty(\mathbb{R})$. This weak convergence follows from the fact that the collection of functions $\{1_{iT}\{y; 0\} - 1_{iT}\{y; 1\} - 1_{iT}\{1\} : y \in \mathbb{R}\}$ is a Vapnik–Çervonenkis class. □

B.4. **Proof of Theorem 2.** Note that

$$\sqrt{n}\big|\sup_{y\in\mathbb{R}}\widehat{\Delta}(y) - \sup_{y\in\mathbb{R}}\Delta(y)\big| \leq \sqrt{n}\sup_{y\in\mathbb{R}}\big|\widehat{\Delta}(y) - \Delta(y)\big| = O_p(1).$$

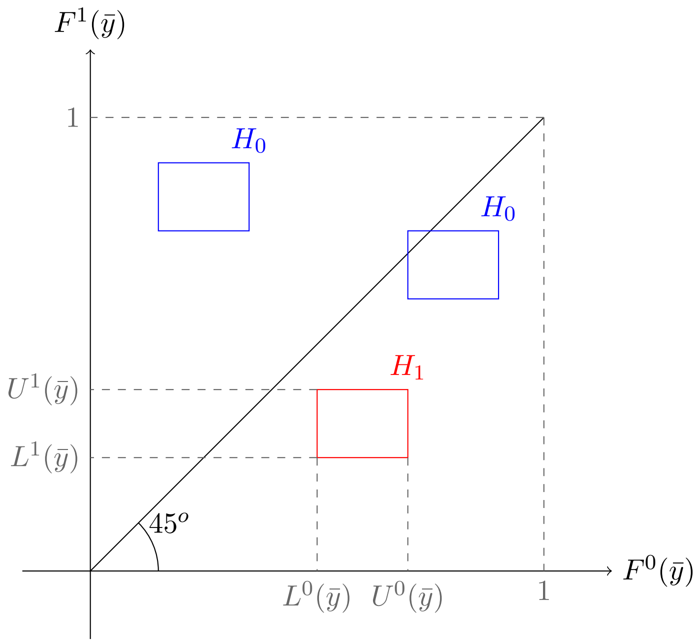


FIGURE 4. Graphical Illustration of the Null and the Alternative

Therefore,

$$\boldsymbol{T}^{KS} = \sqrt{n}\left(\sup_{y\in\mathbb{R}}\widehat{\Delta}(y) - \sup_{y\in\mathbb{R}}\Delta(y)\right) + \sqrt{n}\sup_{y\in\mathbb{R}}\Delta(y) = O_p(1) + \sqrt{n}\sup_{y\in\mathbb{R}}\Delta(y) \to \infty.$$

For $\boldsymbol{T}^{CV}$, similarly

$$\left|\int \max\{\sqrt{n}\left(\widehat{\Delta}(y) - \Delta(y) + \Delta(y)\right),\ 0\}w(y)dy - \int \max\{\sqrt{n}\Delta(y),\ 0\}w(y)dy\right|$$
$$\leq \left|\int \max\{\sqrt{n}\left(\widehat{\Delta}(y) - \Delta(y)\right),\ 0\}w(y)dy\right| = O_p(1),$$

where we used the fact that $\max(a+b,0) \leq \max(a,0) + \max(b,0)$. Therefore,

$$\boldsymbol{T}^{CV} = \int \max\{\sqrt{n}\widehat{\Delta}(y),\ 0\}w(y)dy - \int \max\{\sqrt{n}\Delta(y),\ 0\}w(y)dy$$
$$+ \int \max\{\sqrt{n}\Delta(y),\ 0\}w(y)dy = O_p(1) + \sqrt{n}\int \max\{\Delta(y),\ 0\}w(y)dy \to \infty.$$

□

B.5. **Proof of Theorem 3.** Note that the statistics $\boldsymbol{T}^{KS}$ and $\boldsymbol{T}^{CV}$ are equal to

$$\sup_{y\in\mathbb{R}}\{\sqrt{n}\left(\widehat{\Delta}(y)-\Delta(y)\right)+\delta(y)\}\xrightarrow{d}\sup_{y\in\mathbb{R}}\{\mathbb{G}(y)+\delta(y)\},$$
$$\int_{\mathbb{R}}\max\{\sqrt{n}\left(\widehat{\Delta}(y)-\Delta(y)\right)+\delta(y),\ 0\}w(y)dy\xrightarrow{d}\int_{\mathbb{R}}\max\{\mathbb{G}(y)+\delta(y),\ 0\}w(w)dy,$$

which are simply right–shifted versions of the null distributions of $\boldsymbol{T}^{KS}$ and $\boldsymbol{T}^{CV}$, respectively. As long as $\delta(y)>0$ for some $y\in\mathbb{R}$, the right–shift is strict, which implies that the rejection probability will be strictly larger than the nominal size. □

B.6. **Proof or Theorem 4.** Under H_{0S},

$$\int_{y_{\min}}^{y}\sqrt{n}\widehat{\Delta}(t)dt\le\int_{y_{\min}}^{y}\sqrt{n}\left(\widehat{\Delta}(t)-\Delta(t)\right)dt,$$
$$\max\{\int_{y_{\min}}^{y}\sqrt{n}\widehat{\Delta}(t)dt,\ 0\}\le\max\{\int_{y_{\min}}^{y}\sqrt{n}\left(\widehat{\Delta}(t)-\Delta(t)\right)dt,\ 0\}.$$

Now, the result follows from the continuous mapping theorem and the weak convergence result of Theorem 1. □

B.7. **Proof of Theorem 5.** We first state a set of technical assumptions needed for Theorem 5.

Assumption 6.
- (i) $\widetilde{w}(\cdot,\cdot)$ *is a continuous function on a compact support* $\mathcal{Y}\times\mathcal{X}_s$*, where* $\mathcal{Y}$ *is a strict subset of* $\mathbb{R}$ *and* $\mathcal{X}_s$ *is a strict subset of* $\mathcal{X}$*;*
- (ii) ρ_v *is finite and bounded away from zero on* $\mathcal{Y}\times\mathcal{X}_s$*;*
- (iii) $\bar{\rho}(y,\tilde{y},x)=\rho(y,\tilde{y},x,t)K_*(0)/K_*(t)$ *satisfies* $\bar{\rho}(y,\tilde{y},x)=1-c_1(x)|y-\tilde{y}|^a+o(|y-\tilde{y}|^a)$ *uniformly in* $x\in\mathcal{X}_s$ *as* $|y-\tilde{y}|\to 0$ *for some positive constants* $c_1(x)$ *and* a *such that* c_1 *is bounded away from zero on* $\mathcal{X}_s$*;*
- (iv) K *is an* s*–order kernel with support* $\{u\in\mathbb{R}^{d_T}:\|u\|\le 1/2\}$*, symmetric around zero, integrates to one, and is* s*–times continuously differentiable, where* s *satisfies* $s>3d_T/2$*;*
- (v) *As functions of* x*,* $F^0(y|x),F^1(y|x),f(x),$ *and* $\mathbb{P}[X_i=x|X_i=x]$ *are* s*–times continuously differentiable for each* y *and* x *with uniformly bounded derivatives;*
- (vi) $\sup_{x\in\mathcal{X}_s}F^j(y|x)<\infty$ *for* $j=0,1$*;*
- (vii) *The bandwidth satisfies* $nh_n^{2s}\to 0$*,* $nh_n^{3d_T}\to\infty$*, and* $nh_n^{2d_T}/(\log n)^2\to\infty$*, where* $s>3d_T/2$*.*

Assumption 6 is a modification of Assumption 4.1 of Lee and Whang (2009). For detailed discussions of each condition, check the remarks therein. Note that Part (ii) includes the requirement that f is bounded away from zero on $\mathcal{X}_s$ and that $\bar{\rho}$ in Part (iii) is a correlation coefficient conditional on $X_i=x$, which is equal to one when $y=\tilde{y}$.

(i) Asymptotic Normality: Note first that $\Delta(y,x) = \mathbb{E}[q_i(y)|X_i = x]$. Let

$$\widetilde{\Delta}(y,x) = \frac{1}{f(x)}\frac{1}{n}\sum_{i=1}^{n} q_i(y)K_h(x-X_i) \quad \text{and} \quad R_n(y,x) = \frac{f(x)-\hat{f}(x)}{f(x)\hat{f}(x)}\frac{1}{n}\sum_{i=1}^{n} q_i(y)K_h(x-X_i),$$

which implies that $\widehat{\Delta}(y,x) = \widetilde{\Delta}(y,x) + R_n(y,x)$. Let

$$\zeta_n(y,x) = \Delta(y,x) - \Delta(y,x)\frac{\hat{f}(x)}{f(x)} \quad \text{and}$$
$$\widehat{\Delta}^*(y,x) = \widetilde{\Delta}(y,x) + \zeta_n(y,x).$$

Then, Lemma A.1 of Lee and Whang (2009) shows that $\zeta_n(y,x)$ approximates $R_n(y,x)$ uniformly over $(y,x) \in \mathcal{Y} \times \mathcal{X}_s$ at a rate faster than $n^{1/2}$. Therefore,

$$\widehat{\Delta}^*(y|x) = \widehat{\Delta}(y|x) + o_p\left(n^{-1/2}\right).$$

Define

$$\boldsymbol{T_n}^* = \int\int \sqrt{n}\max\left\{\Delta(y|x) + \widehat{\Delta}^*(y|x) - \mathbb{E}[\widehat{\Delta}^*(y|x)],\ 0\right\}\tilde{w}(y,x)dydx. \tag{13}$$

Then, Lemma A.2 of Lee and Whang (2009) shows that $\boldsymbol{T}^C = \boldsymbol{T_n}^* + o_p(1)$. Since $\Delta(y|x) = 0$ for all $(y,x) \in \mathcal{Y} \times \mathcal{X}_s$ under the least favorable null hypothesis, we plug $\Delta(y|x) = 0$ into $\boldsymbol{T_n}^*$ and define $\boldsymbol{T_n}$ as follows:

$$\boldsymbol{T_n} = \int\int \sqrt{n}\max\left\{\widehat{\Delta}^*(y|x) - \mathbb{E}\left[\widehat{\Delta}^*(y|x)\right],\ 0\right\}\tilde{w}(y,x)dydx.$$

Then, it is enough to show that

$$\frac{\boldsymbol{T_n} - \eta_n}{\sigma_0} \xrightarrow{d} N(0,1),$$

which follows from Theorem A.2 in Lee and Whang (2009).

(ii) Consistency: This result comes directly from the proof of Theorem 4.2 of Lee and Whang (2009) by changing $\tau_0(y,x)$ and $w(y,x)$ with $\Delta(y,x)$ and $\widetilde{w}(y,x)$.

(iii) Local asymptotic power: We have

$$\frac{\widehat{\boldsymbol{T}}^C - \eta_n}{\sigma_0} = \frac{\widehat{\boldsymbol{T}}^C - \widetilde{\eta}}{\sigma_0} + \frac{\widetilde{\eta} - \eta_n}{\sigma_0}.$$

where

$$\widetilde{\eta} = \int\int \mathbb{E}\max\left\{\delta(y,x) + h^{-d_T/2}\sqrt{\rho_v(y,x)}\mathbb{Z}_1, 0\right\}\widetilde{w}(y,x)\,dydx.$$

From Theorem 4.3 in Lee and Whang (2009), we know that, under the local alternatives $H_{1C,n}$,

$$\frac{\widehat{\boldsymbol{T}}^C - \widetilde{\eta}}{\sigma_0} \xrightarrow{d} N(0,1).$$

Now, note that

$$\sigma_0 \left(\frac{\widetilde{\eta} - \eta_n}{\sigma_0} \right) \geq$$

$$\int \int \mathbb{E} \max \left\{ \delta (y, x) + h^{-d_T/2} \sqrt{\rho_v (y, x)} \mathbb{Z}_1, 0 \right\} - \max \left\{ h^{-d_T/2} \sqrt{\rho_v (y, x)} \mathbb{Z}_1, 0 \right\} w (y, x) \, dy dx$$

$$\geq \frac{1}{2} \int \int \delta (y, x) \, \widetilde{w} (y, x) \, dy dx > 0,$$

which establishes the nontrivial local power property. □

Appendix C. A Remark on Computation

In Section 3, test statistics are obtained by averaging over the $T!$ permutations. In practice, however, we do not need to consider the entire permutations when calculating $1_{iT}\{y; j\}$ and $1_{iT}\{j\}$, because (i) most of the indicators are zero and (ii) many nonzero indicators are the same.

Note that, for each i, the treatment history $D_i = (D_{i1}, \cdots, D_{iT})'$ is given as one of the 2^T types of $T \times 1$ indicator vectors. For example, when $T = 2$, person i could have one of the following four types of treatment history D_i: $(0, 0)'$, $(0, 1)'$, $(1, 0)'$ or $(1, 1)'$. Therefore, when $j = 1$, $1_i\{y; 1, k\}$ is possibly nonzero only for the following cases:

- When $D_i = (1, 1)'$, $1_i\{y; 1, k\} = 1\{Y_{i1} \leq y, D_{i1} = 1, D_{i2} = 1\}$ for $(t_1, t_2) = (1, 2)$ permutation; or $1_i\{y; 1, k\} = 1\{Y_{i2} \leq y, D_{i1} = 1, D_{i2} = 1\}$ for $(t_1, t_2) = (2, 1)$.
- When $D_i = (1, 0)'$, $1_i\{y; 1, k\} = 1\{Y_{i1} \leq y, D_{i1} = 1, D_{i2} = 0\}$ for either $(t_1, t_2) = (1, 2)$ or $(2, 1)$.
- When $D_i = (0, 1)'$, $1_i\{y; 1, k\} = 1\{Y_{i2} \leq y, D_{i1} = 0, D_{i2} = 1\}$ for either $(t_1, t_2) = (1, 2)$ or $(2, 1)$.
- When $D_i = (0, 0)'$, there exist no nonzero cases.

Over the permutations, using the same argument, we can conclude that $1_i\{y; j, k\} = 1$ only when $1\{Y_{it} \leq y, D_{it} = j\} = 1$ for $t = 1, 2, \cdots, T$ (i.e., when the time index of Y_{it} corresponds to that of D_{it}). It implies that $1_{iT}\{y; j\}$ can be simply obtained as $1_{iT}\{y; j\} = |D_i|_j^{-1} \sum_{t=1}^{T} 1\{Y_{it} \leq y, D_{it} = j\}$ for $j = 0, 1$, where $|D_i|_j$ counts the number of elements $D_{it} = j$ in D_i. Similarly, $1_{iT}\{j\} = 1\{D_{i1} = 1 - j, \cdots, D_{iT} = 1 - j\}$. This remark tells that the computational cost of estimating the test statistics is very minimal since we do not need to go over the permutations in practice.

References Cited

Abadie, A. (2002): "Bootstrap tests for distributional treatment effects in instrumental variable models," *Journal of the American Statistical Association*, 97(457), 284–292.

Abrevaya, J. (2006): "Estimating the effect of smoking on birth outcomes using a matched panel data approach," *Journal of Applied Econometrics*, 21(4), 489–519.

Abrevaya, J., and C. M. Dahl (2008): "The Effects of Birth Inputs on Birthweight: Evidence From Quantile Estimation on Panel Data," *Journal of Business and Economic Statistics*, 155, 56–70.

Andrews, D. W. K., and G. Soares (2010): "Inference for Parameters Defined by Moment Inequalities Using Generalized Moment Selection," *Econometrica*, 78(1), 119–157.

Angrist, J., G. Imbens, and D. Rubin (1996): "Identification of causal effects using instrumental variables.," *Journal of the American Statistical Association*, 91(434).

Arellano, M., and S. Bonhomme (2010): "Identifying Distributional Characteristics in Random Coefficients Panel Data Models," Discussion paper, CEMFI.

Ashenfelter, O., and D. Card (1985): "Using the longitudinal structure of earnings to estimate the effect of training programs," *The Review of Economics and Statistics*, 67(4), 648–660.

Barrett, G., and S. Donald (2003): "Consistent tests for stochastic dominance," *Econometrica*, 71(1), 71–104.

Beck, M., and B. Fitzenberger (2004): "Changes in Union Membership Over Time: A Panel Analysis for West Germany," *LABOUR*, 18, 329–362.

Beresteanu, A., and F. Molinari (2008): "Asymptotic Properties for a Class of Partially Identified Models," *Econometrica*, 76(4), 763–814.

Blakemore, A., J. Hunt, and B. Kiker (1986): "Collective bargaining and union membership effects on the wages of male youths," *Journal of Labor Economics*, 4(2), 193–211.

Budd, J., and I. Na (2000): "The union membership wage premium for employees covered by collective bargaining agreements," *Journal of Labor Economics*, 18(4), 783–807.

Chernozhukov, V., I. Fernandez-Val, J. Hahn, and W. Newey (2009): "Identification and estimation of marginal effects in nonlinear panel models," Discussion paper, CEMMAP.

Chernozhukov, V., I. Fernandez-Val, and W. Newey (2009): "Quantile and average effects in nonseparable panel models," Discussion paper, CEMMAP.

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

Crump, R. K., V. J. Hotz, G. W. Imbens, and O. A. Mitnik (2008): "Nonparametric Tests for Treatment Effect Heterogeneity," *The Review of Economics and Statistics*, 90(3), 389–405.

Dehejia, R., and S. Wahba (2002): "Propensity score-matching methods for nonexperimental causal studies," *Review of Economics and statistics*, 84(1), 151–161.

Djebbari, H., and J. Smith (2008): "Heterogeneous impacts in PROGRESA," *Journal of Econometrics*, 145(1-2), 64–80.

Evans, W., and J. S. Ringel (1999): "Can Higher Cigarette Taxes Improve Birth Outcomes?," *Journal of Public Economics*, 72, 135–154.

Firpo, S. (2007): "Efficient semiparametric estimation of quantile treatment effects," *Econometrica*, 75(1), 259–276.

Giné, E., and J. Zinn (1990): "Bootstrapping general empirical measures," *The Annals of Probability*, 18(2), 851–869.

Graham, B. S., and J. Powell (2008): "Identification and Estimation of 'Irregular' Correlated Random Coefficient Models," NBER Working Papers 14469, National Bureau of Economic Research, Inc.

Hahn, J. (1998): "On the role of the propensity score in semiparametric estimation of average treatment effects," *Econometrica*, 66, 315–331.

Hahn, J., and G. Ridder (2009): "Partial identification and confidence intervals," Discussion paper, UCLA.

Hansen, B. (1996): "Inference when a nuisance parameter is not identified under the null hypothesis," *Econometrica: Journal of the Econometric Society*, pp. 413–430.

Heckman, J., and J. Smith (1997): "Making the Most Out Of Programme Evaluations and Social Experiments: Accounting For Heterogeneity in Programme Impacts," *Review of Economics Studies*, 64, 487–535.

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

——— (2007b): "Econometric Evaluation of Social Programs, Part II: Using the Marginal Treatment Effect to Organize Alternative Econometric Estimators to Evaluate Social Programs, and to Forecast their Effects in New Environments," in *Handbook of Econometrics*, ed. by J. Heckman, and E. Leamer, vol. 6B of *Handbook of Econometrics*, chap. 71, pp. 4875–5143. Elsevier.

Hirano, K., G. Imbens, and G. Ridder (2003): "Efficient estimation of average treatment effects using the estimated propensity score," *Econometrica*, 71, 1161–1189.

Hoderlein, S., and Y. Sasaki (2011): "Continuous Treatments," Discussion paper, Boston College.

IMBENS, G., AND C. MANSKI (2004): "Confidence Intervals for Partially Identified Parameters," *Econometrica*, 74, 1845–1857.

IMBENS, G. W., AND J. M. WOOLDRIDGE (2009): "Recent Developments in the Econometrics of Program Evaluation," *Journal of Economic Literature*, 47(1), 5–86.

JONES, E. (1982): "Union/Nonunion differentials: Membership or coverage?," *Journal of Human Resources*, 17(2), 276–285.

JUNG, K. (2010): "Incentives for Voluntary Disclosure of Quality Information in HMO Markets," *Journal of Risk and Insurance*, 77(1), 183–210.

KHAN, S., M. PONOMAREVA, AND E. TAMER (2011): "Identification of Panel Data Models with Endogenous Censoring," Discussion paper, Duke University.

LEE, M. J. (2009): "Non-parametric tests for distributional treatment effect for randomly censored responses," *Journal of the Royal Statistical Society Series B*, 71(1), 243–264.

LEE, S., AND Y. J. WHANG (2009): "Nonparametric Tests of Conditional Treatment Effects," Discussion paper, Cowles Foundation.

LEMIEUX, T. (1998): "Estimating the Effects of Unions on Wage Inequality in a Panel Data Model with Comparative Advantage and Nonrandom Selection," *Journal of Labor Economics*, 16(2), 261–291.

MANSKI, C. F. (1990): "Nonparametric Bounds on Treatment Effects," *American Economic Review*, 80(2), 319–323.

MATZKIN, R. L. (2007): "Nonparametric Identification," in *Handbook of Econometrics*, ed. by J. Heckman, and E. Leamer, vol. 6B of *Handbook of Econometrics*, chap. 73, pp. 5307–5368. Elsevier.

PERMUTT, T., AND J. HEBEL (1989): "Simultaneous-equation estimation in a clinical trial of the effect of smoking on birth weight," *Biometrics*, 45(2), 619–622.

ROBINSON, C. (1989): "The Joint Determination of Union Status and Union Wage Effects: Some Tests of Alternative Models," *Journal of Political Economy*, 97(3), 639–667.

ROMANO, J. P., AND A. M. SHAIKH (2010): "Inference for the Identified Set in Partially Identified Econometric Models," *Econometrica*, 78(1), 169–211.

ROSEN, A. M. (2008): "Confidence sets for partially identified parameters that satisfy a finite number of moment inequalities," *Journal of Econometrics*, 146(1), 107–117.

ROSENBAUM, P., AND D. RUBIN (1983): "The central role of the propensity score in observational studies of causal effects," *Biometrika*, 70, 41–55.

ROTHE, C. (2010): "Nonparametric estimation of distributional policy effects," *Journal of Econometrics*, 155, 56–70.

WOOLDRIDGE, J. M. (2005): "Fixed-effects and related estimators for correlated-random coefficient and treatment-effect panel data models," *Review of Economics and Statistics*, 87(2), 385 – 390.