Econometrica, Vol. 74, No. 2 (March, 2006), 431–497

IDENTIFICATION AND INFERENCE IN NONLINEAR DIFFERENCE-IN-DIFFERENCES MODELS

BY SUSAN ATHEY AND GUIDO W. IMBENS[1]

This paper develops a generalization of the widely used difference-in-differences method for evaluating the effects of policy changes. We propose a model that allows the control and treatment groups to have different average benefits from the treatment. The assumptions of the proposed model are invariant to the scaling of the outcome. We provide conditions under which the model is nonparametrically identified and propose an estimator that can be applied using either repeated cross section or panel data. Our approach provides an estimate of the entire counterfactual distribution of outcomes that would have been experienced by the treatment group in the absence of the treatment and likewise for the untreated group in the presence of the treatment. Thus, it enables the evaluation of policy interventions according to criteria such as a mean–variance trade-off. We also propose methods for inference, showing that our estimator for the average treatment effect is root-N consistent and asymptotically normal. We consider extensions to allow for covariates, discrete dependent variables, and multiple groups and time periods.

KEYWORDS: Difference-in-differences, identification, nonlinear models, heterogenous treatment effects, nonparametric estimation.

1. INTRODUCTION

DIFFERENCE-IN-DIFFERENCES (DID) methods for estimating the effect of policy interventions have become very popular in economics.[2] These methods are used in problems with multiple subpopulations—some subject to a policy intervention or treatment and others not—and outcomes that are measured in each group before and after the policy intervention (although not necessarily for the same individuals).[3] To account for time trends unrelated to the interven-

[1]We are grateful to Alberto Abadie, Joseph Altonji, Don Andrews, Joshua Angrist, David Card, Esther Duflo, Austan Goolsbee, Jinyong Hahn, Caroline Hoxby, Rosa Matzkin, Costas Meghir, Jim Poterba, Scott Stern, Petra Todd, Edward Vytlacil, seminar audiences at the University of Arizona, UC Berkeley, the University of Chicago, University of Miami, Monash University, Harvard/MIT, Northwestern University, UCLA, USC, Yale University, Stanford University, the San Francisco Federal Reserve Bank, the Texas Econometrics conference, SITE, NBER, and AEA 2003 winter meetings, the 2003 Joint Statistical Meetings, and, especially, Jack Porter for helpful discussions. We are indebted to Bruce Meyer, who generously provided us with his data. Four anonymous referees and a co-editor provided insightful comments. Richard Crump, Derek Gurney, Lu Han, Khartik Kalyanaram, Peyron Law, Matthew Osborne, Leonardo Rezende, and Paul Riskind provided skillful research assistance. Financial support for this research was generously provided through NSF grants SES-9983820 and SES-0351500 (Athey), SBR-9818644, and SES 0136789 (Imbens).

[2]In other social sciences such methods are also widely used, often under other labels such as the "untreated control group design with independent pretest and posttest samples" (e.g., Shadish, Cook, and Campbell (2002)).

[3]Examples include the evaluation of labor market programs (Ashenfelter and Card (1985), Blundell, Costa Dias, Meghir, and Van Reenen (2001)), civil rights (Heckman and Payner (1989),

tion, the change experienced by the group subject to the intervention (referred to as the treatment group) is adjusted by the change experienced by the group not subject to treatment (the control group). Several recent surveys describe other applications and give an overview of the methodology, including Meyer (1995), Angrist and Krueger (2000), and Blundell and MaCurdy (2000).

This paper analyzes nonparametric identification, estimation, and inference for the average effect of the treatment for settings where repeated cross sections of individuals are observed in a treatment group and a control group, before and after the treatment. Our approach differs from the standard DID approach in several ways. We allow the effects of both time and the treatment to differ systematically across individuals,[4] as when inequality in the returns to skill increases over time or when new medical technology differentially benefits sicker patients. We propose an estimator for the entire counterfactual distribution of effects of the treatment on the treatment group as well as the distribution of effects of the treatment on the control group, where the two distributions may differ from each other in arbitrary ways. We accommodate the possibility—but do not assume—that the treatment group adopted the policy because it expected greater benefits than the control group. (Besley and Case (2000) discuss this possibility as a concern for standard DID models.) In contrast, standard DID methods give little guidance about what the effect of a policy intervention would be in the (counterfactual) event that it were applied to the control group except in the extreme case where the effect of the policy is constant across individuals.

We develop our approach in several steps. First, we develop a new model that relates outcomes to an individual's group, time, and unobservable characteristics.[5] The standard DID model is a special case of our model, which we call the *changes-in-changes* model. In the standard model, groups and time periods are treated symmetrically: for a particular scaling of the outcomes, the mean of individual outcomes in the absence of the treatment is additive in group and

---

Donohue, Heckman, and Todd (2002)), the inflow of immigrants (Card (1990)), the minimum wage (Card and Krueger (1993)), health insurance (Gruber and Madrian (1994)), 401(k) retirement plans (Poterba, Venti, and Wise (1995)), worker's compensation (Meyer, Viscusi, and Durbin (1995)), tax reform (Eissa and Liebman (1996), Blundell, Duncan, and Meghir (1998)), 911 systems (Athey and Stern (2002)), school construction (Duflo (2001)), information disclosure (Jin and Leslie (2003)), World War II internment camps (Chin (2005)), and speed limits (Ashenfelter and Greenstone (2004)). Time variation is sometimes replaced by another type of variation, as in Borenstein's (1991) study of airline pricing.

[4]Treatment effect heterogeneity has been a focus of the general evaluation literature, e.g., Heckman and Robb (1985), Manski (1990), Imbens and Angrist (1994), Dehejia (1997), Lechner (1999), Abadie, Angrist, and Imbens (2002), and Chernozhukov and Hansen (2005), although it has received less attention in difference-in-differences settings.

[5]The proposed model is related to models of wage determination proposed in the literature on wage decomposition where changes in the wage distribution are decomposed into changes in returns to (unobserved) skills and changes in relative skill distributions (Juhn, Murphy, and Pierce (1991), Altonji and Blank (2000)).

time indicators.[6] In contrast, in our model, time periods and groups are treated asymmetrically. The defining feature of a time period is that in the absence of the treatment, within a period the outcomes for all individuals are determined by a single, monotone "production function" that maps individual-specific unobservables to outcomes. The defining feature of a group is that the distribution of individual unobservable characteristics is the same within a group in both time periods, even though the characteristics of any particular agent can change over time. Groups can differ in arbitrary ways in the distribution of the unobserved individual characteristic and, in particular, the treatment group might have more individuals who experience a high return to the treatment.

Second, we provide conditions under which the proposed model is identified nonparametrically and we develop a novel estimation strategy based on the identification result. We use the entire "before" and "after" outcome distributions in the control group to nonparametrically estimate the change over time that occurred in the control group. Assuming that the distribution of outcomes in the treatment group would have experienced the same change in the absence of the intervention, we estimate the counterfactual distribution for the treatment group in the second period. We compare this counterfactual distribution to the actual second-period distribution for the treatment group. Thus, we can estimate—without changing the assumptions underlying the estimators—the effect of the intervention on any feature of the distribution. We use a similar approach to estimate the effect of the treatment on the control group.

A third contribution is to develop the asymptotic properties of our estimator. Estimating the average and quantile treatment effects involves estimating the inverse of an empirical distribution function with observations from one group–period and applying that function to observations from a second group–period (and averaging this transformation for the average treatment effect). We establish root-N consistency and asymptotic normality of the estimator for the average treatment effect and quantile treatment effects. We extend the analysis to incorporate covariates.

In a fourth contribution, we extend the model to allow for discrete outcomes. With discrete outcomes, the standard DID model can lead to predictions outside the allowable range. These concerns have led researchers to consider nonlinear transformations of an additive single index. However, the economic justification for the additivity assumptions required for DID may be tenuous in such cases. Because we do not make functional form assumptions, this problem does not arise using our approach. However, without additional assumptions, the counterfactual distribution of outcomes may not be identified when outcomes are discrete. We provide bounds (in the spirit of Manski (1990, 1995))

[6]We use the term "standard DID model" to refer to a model that assumes that outcomes are additive in a time effect, a group effect, and an unobservable that is independent of the time and group (e.g., Meyer (1995), Angrist and Krueger (2000), and Blundell and MaCurdy (2000)). The scale-dependent additivity assumptions of this model have been criticized as unduly restrictive from an economic perspective (e.g., Heckman (1996)).

on the counterfactual distribution and show that the bounds collapse as the outcomes become "more continuous." We then discuss two alternative approaches for restoring point identification. The first alternative relies on an additional assumption about the unobservables. It leads to an estimator that differs from the standard DID estimator even for the simple binary response model without covariates. The second alternative is based on covariates that are independent of the unobservable. Such covariates can tighten the bounds or even restore point identification.

Fifth, we consider an alternative approach to constructing the counterfactual distribution of outcomes in the absence of treatment—the "quantile DID" (QDID) approach. In the QDID approach we compute the counterfactual distribution by adding the change over time at the qth quantile of the control group to the qth quantile of the first-period treatment group. Meyer, Viscusi, and Durbin (1995) and Poterba, Venti, and Wise (1995) apply this approach to specific quantiles. We propose a nonlinear model for outcomes that justifies the quantile DID approach for every quantile simultaneously and thus validates construction of the entire counterfactual distribution. The standard DID model is a special case of this model. Despite the intuitive appeal of the quantile DID approach, we show that the underlying model has several unattractive features.

Sixth, we provide extensions to settings with multiple groups and multiple time periods.

Finally, in the supplementary material to this article, available on the *Econometrica* website (Athey and Imbens (2006)), we apply the methods developed in this paper to study the effects of disability insurance on injury durations using data previously analyzed by Meyer, Viscusi, and Durbin (1995). This application shows that the approach used to estimate the effects of a policy change can lead to results that differ from the standard DID results in terms of magnitude and significance. Thus, the restrictive assumptions required for standard DID methods can have significant policy implications. We also present simulations that illustrate the small sample properties of the estimators and highlight the potential importance of accounting for the discrete nature of the data.

Some of the results developed in this paper can also be applied outside of the DID setting. For example, our estimator for the average treatment effect for the treated is closely related to an estimator proposed by Juhn, Murphy, and Pierce (1991) and Altonji and Blank (2000) to decompose the Black–White wage differential into changes in the returns to skills and changes in the relative skill distribution.[7] As we discuss below, our asymptotic results apply to the Altonji–Blank estimator and, furthermore, our results for discrete data extend their model.

Within the literature on treatment effects, the results in this paper are most closely related to the literature concerning panel data. In contrast, our ap-

[7]See also the work by Fortin and Lemieux (1999) on the gender gap in wage distributions.

proach is tailored to the case of repeated cross sections. A few recent papers analyze the theory of DID models, but their focus differs from ours. Abadie (2005) and Blundell, Costa Dias, Meghir, and Van Reenen (2001) discuss adjusting for exogenous covariates using propensity score methods. Donald and Lang (2001) and Bertrand, Duflo, and Mullainathan (2004) address problems with standard methods for computing standard errors in DID models; their solutions require multiple groups and periods, and rely heavily on linearity and additivity.

Finally, we note that our approach to nonparametric identification relies heavily on an assumption that in each time period, the "production function" is monotone in an unobservable. Following Matzkin (1999, 2003), Altonji and Matzkin (1997, 2005), and Imbens and Newey (2001), a growing literature exploits monotonicity in the analysis of nonparametric identification of nonseparable models; we discuss this literature in more detail below.

2. GENERALIZING THE STANDARD DID MODEL

The standard model for the DID design is as follows. Individual i belongs to a group $G_i \in \{0, 1\}$ (where group 1 is the treatment group) and is observed in time period $T_i \in \{0, 1\}$. For $i = 1, \ldots, N$, a random sample from the population, individual i's group identity and time period can be treated as random variables. Letting the outcome be Y_i, the observed data are the triple (Y_i, G_i, T_i).[8] Using the potential outcome notation advocated in the treatment effect literature by Rubin (1974, 1978), let Y_i^N denote the outcome for individual i if that individual does not receive the treatment, and let Y_i^I be the outcome for the same individual if he or she does receive the treatment. Thus, if I_i is an indicator for the treatment, the realized (observed) outcome for individual i is

$$Y_i = Y_i^N \cdot (1 - I_i) + I_i \cdot Y_i^I.$$

In the two-group–two-period setting we consider, $I_i = G_i \cdot T_i$.

In the standard DID model, the outcome for individual i in the absence of the intervention satisfies

$$Y_i^N = \alpha + \beta \cdot T_i + \gamma \cdot G_i + \varepsilon_i. \tag{1}$$

The second coefficient, β, represents the time effect. The third coefficient, γ, represents a group-specific time-invariant effect.[9] The third term, ε_i, represents unobservable characteristics of the individual. This term is assumed to be

[8] In Sections 4 and 5 and we discuss cases with exogenous covariates.

[9] In some settings, it is more appropriate to generalize the model to allow for a time-invariant individual-specific fixed effect γ_i, potentially correlated with G_i. See, e.g., Angrist and Krueger (2000). This generalization of the standard model does not affect the standard DID estimand and it will be subsumed as a special case of the model we propose. See Section 3.4 for more discussion of panel data.

independent of the group indicator and have the same distribution over time, i.e., $\varepsilon_i \perp (G_i, T_i)$, and is normalized to have mean zero. The standard DID estimand is

$$\tau^{\text{DID}} = \big[\mathbb{E}[Y_i|G_i=1, T_i=1] - \mathbb{E}[Y_i|G_i=1, T_i=0]\big] - \big[\mathbb{E}[Y_i|G_i=0, T_i=1] - \mathbb{E}[Y_i|G_i=0, T_i=0]\big]. \tag{2}$$

In other words, the population average difference over time in the control group ($G_i = 0$) is subtracted from the population average difference over time in the treatment group ($G_i = 1$) to remove biases associated with a common time trend unrelated to the intervention.

Note that the full independence assumption $\varepsilon_i \perp (G_i, T_i)$ (e.g., Blundell and MaCurdy (2000)) is stronger than necessary for τ^{DID} to give the average treatment effect. One can generalize this framework and allow for general forms of heteroskedasticity by group or time by relaxing the assumption to only mean independence (e.g., Abadie (2005)) or zero correlation between ε_i and (G_i, T_i). Our proposed model will nest the DID model with independence (which for further reference will be labeled the standard DID model), but not the DID model with mean independence.[10]

The interpretation of the standard DID estimand depends on assumptions about how outcomes are generated in the presence of the intervention. It is often assumed that the treatment effect is constant across individuals, so that $Y_i^I - Y_i^N = \tau$. Combining this restriction with the standard DID model for the outcome without intervention leads to a model for the realized outcome:

$$Y_i = \alpha + \beta \cdot T_i + \gamma \cdot G_i + \tau \cdot I_i + \varepsilon_i.$$

More generally, the effect of the intervention might differ across individuals. Then the standard DID estimand gives the average effect of the intervention on the treatment group.

We propose to generalize the standard model in several ways. First, we assume that in the absence of the intervention, the outcomes satisfy

$$Y_i^N = h(U_i, T_i), \tag{3}$$

with $h(u, t)$ increasing in u. The random variable U_i represents the unobservable characteristics of individual i, and (3) incorporates the idea that the outcome of an individual with $U_i = u$ will be the same in a given time period,

[10]The DID model with mean independence assumes that, for a given scaling of the outcome, changes across subpopulations in the mean of Y_i have a structural interpretation and as such are used to predict the counterfactual outcome for the second-period treatment group in the absence of the treatment. In contrast, all differences across subpopulations in the other moments of the distribution of Y_i are ignored when making predictions. In the model we propose, all changes in the distribution of Y_i across subpopulations are given a structural interpretation and used for inference. Neither our model nor the DID model with mean independence imposes any testable restrictions on the data.

irrespective of the group membership. The distribution of U_i is allowed to vary across groups, but not over time within groups, so that $U_i \perp T_i|G_i$. The standard DID model in (1) embodies three additional assumptions, namely

$$\text{(4)} \qquad \text{(additivity)} \quad U_i = \alpha + \gamma \cdot G_i + \varepsilon_i \quad \text{with} \quad \varepsilon_i \perp (G_i, T_i),$$

$$\text{(5)} \qquad \text{(single index model)} \quad h(u,t) = \phi(u + \delta \cdot t)$$

for a strictly increasing function $\phi(\cdot)$, and

$$\text{(6)} \qquad \text{(identity transformation)} \quad \phi(\cdot) \text{ is the identity function.}$$

Thus the proposed model nests the standard DID as a special case. The mean-independence DID model is not nested; rather, the latter model requires that changes over time in moments of the outcomes other than the mean are not relevant for predicting the mean of Y_i^N. Note also that in contrast to the standard DID model, our assumptions do not depend on the scaling of the outcome, for example, whether outcomes are measured in levels or logarithms.[11]

A natural extension of the standard DID model might have been to maintain assumptions (4) and (5) but relax (6), to allow $\phi(\cdot)$ to be an unknown function.[12] Doing so would maintain an additive single index structure within an unknown transformation, so that

$$\text{(7)} \qquad Y_i^N = \phi(\alpha + \gamma \cdot G_i + \delta \cdot T_i + \varepsilon_i).$$

However, this specification still imposes substantive restrictions, for example, ruling out some models with mean and variance shifts both across groups and over time.[13]

In the proposed model, the treatment group's distribution of unobservables may be different from that of the control group in arbitrary ways. In the absence of treatment, *all* differences between the two groups can be attributed to differences in the conditional distribution of U given G. The model further requires that the changes over time in the distribution of each group's outcome (in the absence of treatment) arise from the fact that $h(u,0)$ differs from $h(u,1)$, that is, the effect of the unobservable on outcomes changes over time. Like the standard model, our approach does not rely on tracking individuals over time; although the distribution of U_i is assumed not to change over

[11]To be precise, we say that a model is invariant to the scaling of the outcome if, given the validity of the model for Y, the same assumptions remain valid for any strictly monotone transformation of the outcome.

[12]Ashenfelter and Greenstone (2004) consider models where $\phi(\cdot)$ is a Box–Cox transformation with unknown parameter.

[13]For example, suppose that $Y_i^N = \alpha + \delta_1 \cdot T_i + (\gamma \cdot G_i + \varepsilon_i) \cdot (1 + \delta_2 \cdot T_i)$. In the second period there is a shift in the mean as well as an unrelated shift in the variance, meaning the model is incompatible with (7).

time within groups, we do not make any assumptions about whether a particular individual has the same realization U_i in each period. Thus, the estimators we derive for our model will be the same whether we observe a panel of individuals over time or a repeated cross section. We discuss alternative models for panel data in more detail in Section 3.4.

Just as in the standard DID approach, if we wish to estimate only the effect of the intervention on the treatment group, no assumptions are required about how the intervention affects outcomes. To analyze the counterfactual effect of the intervention on the control group, we assume that in the presence of the intervention,

$$Y_i^I = h^I(U_i, T_i)$$

for some function $h^I(u, t)$ that is increasing in u. That is, the effect of the treatment at a given time is the same for individuals with the same $U_i = u$, irrespective of the group. No further assumptions are required on the functional form of h^I, so the treatment effect, equal to $h^I(u, 1) - h(u, 1)$ for individuals with unobserved component u, can differ across individuals. Because the distribution of the unobserved component U can vary across groups, the average return to the policy intervention can vary across groups as well.

3. IDENTIFICATION IN MODELS WITH CONTINUOUS OUTCOMES

3.1. *The Changes-in-Changes Model*

This section considers identification of the changes-in-changes (CIC) model. We modify the notation by dropping the subscript i and treating (Y, G, T, U) as a vector of random variables. To ease the notational burden, we introduce the shorthand

$$Y_{gt}^N \overset{d}{\sim} Y^N | G = g, T = t, \quad Y_{gt}^I \overset{d}{\sim} Y^I | G = g, T = t,$$

$$Y_{gt} \overset{d}{\sim} Y | G = g, T = t, \quad U_g \overset{d}{\sim} U | G = g,$$

where $\overset{d}{\sim}$ is shorthand for "is distributed as." The corresponding conditional distribution functions are $F_{Y^N,gt}$, $F_{Y^I,gt}$, $F_{Y,gt}$, and $F_{U,g}$, with supports $\mathbb{Y}_{gt}^N$, $\mathbb{Y}_{gt}^I$, $\mathbb{Y}_{gt}$, and $\mathbb{U}_g$, respectively.

We analyze sets of assumptions that identify the distribution of the counterfactual second-period outcome for the treatment group, that is, sets of assumptions that allow us to express the distribution $F_{Y^N,11}$ in terms of the joint distribution of the observables (Y, G, T). In practice, these results allow us to express $F_{Y^N,11}$ in terms of the three estimable conditional outcome distributions in the other three subpopulations not subject to the intervention, $F_{Y,00}$, $F_{Y,01}$, and $F_{Y,10}$. Consider first a model of outcomes in the absence of the intervention.

ASSUMPTION 3.1—Model: *The outcome of an individual in the absence of intervention satisfies the relationship* $Y^N = h(U, T)$.

The next set of assumptions restricts h and the joint distribution of (U, G, T).

ASSUMPTION 3.2—Strict Monotonicity: *The production function* $h(u, t)$, *where* $h : \mathbb{U} \times \{0, 1\} \to \mathbb{R}$, *is strictly increasing in* u *for* $t \in \{0, 1\}$.

ASSUMPTION 3.3—Time Invariance Within Groups: *We have* $U \perp T|G$.

ASSUMPTION 3.4—Support: *We have* $\mathbb{U}_1 \subseteq \mathbb{U}_0$.

Assumptions 3.1–3.3 comprise the CIC model; we will invoke Assumption 3.4 selectively for some of the identification results as needed. When the outcomes are continuous, the assumptions of the CIC model (Assumptions 3.1–3.3) do not restrict the data and thus the model is not testable.

Assumption 3.1 requires that outcomes do not depend directly on the group indicator and further that all relevant unobservables can be captured in a single index, U. The assumption of a single index can be restrictive. If $h(u, t)$ is nonlinear, this assumption rules out, for example, the presence of classical measurement error on the outcome. Assumption 3.2 requires that higher unobservables correspond to strictly higher outcomes. Such monotonicity arises naturally when the unobservable is interpreted as an individual characteristic such as health or ability. Within a single time period, monotonicity is simply a normalization, but requiring monotonicity in both periods places restrictions on the way the production function changes over time. Strict monotonicity is automatically satisfied in additively separable models, but it allows for a rich set of nonadditive structures that arise naturally in economic models. The distinction between strict and weak monotonicity is innocuous in models where the outcomes Y_{gt} are continuous.[14] However, in models where there are mass points in the distribution of Y_{gt}^N, strict monotonicity is unnecessarily restrictive.[15] In Section 4, we focus specifically on discrete outcomes and relax this assumption; the results in this section are intended primarily for models with continuous outcomes.

Assumption 3.3 requires that the population of agents within a given group does not change over time. This strong assumption is at the heart of both the DID and CIC approaches. It requires that any differences between the groups be stable, so that estimating the trend on one group can assist in eliminating the

[14]To see this, observe that if Y_{gt} is continuous and h is nondecreasing in u, Y_{gt} and U_g must be one-to-one, and so U_g is continuous as well. However, then h must be strictly increasing in u.

[15]Whereas $Y_{gt} = h(U_g, t)$, strict monotonicity of h implies that each mass point of Y_{g0} corresponds to a mass point of equal size in the distribution of Y_{g1}.

trend in the other group. Under this assumption, any change in the variance of outcomes over time within a group will be attributed to changes over time in the production function. In contrast, the standard DID model with full independence rules out such changes and the DID model with mean independence ignores such changes. Assumption 3.4 implies that $\mathbb{Y}_{10} \subseteq \mathbb{Y}_{00}$ and $\mathbb{Y}_{11}^N \subseteq \mathbb{Y}_{01}$; we relax this assumption in a corollary of the identification theorem.

Our analysis makes heavy use of inverse distribution functions. We will use the convention that, for $q \in [0, 1]$ and for a random variable Y with compact support $\mathbb{Y}$,

$$F_Y^{-1}(q) = \inf\{y \in \mathbb{Y} : F_Y(y) \geq q\}. \tag{8}$$

This implies that the inverse distribution functions are continuous from the left and, for all $q \in [0, 1]$, we have $F_Y(F_Y^{-1}(q)) \geq q$ with equality at all $y \in \mathbb{Y}$ for continuous Y and at discontinuity points of $F_Y^{-1}(q)$ for discrete Y. In addition, $F_Y^{-1}(F_Y(y)) \leq y$, again with equality at all $y \in \mathbb{Y}$ for continuous or discrete Y, but not necessarily if Y is mixed.

Identification for the CIC model is established in the following theorem.

THEOREM 3.1—Identification of the CIC Model: *Suppose that Assumptions* 3.1–3.4 *hold, and that U is either continuous or discrete. Then the distribution of* Y_{11}^N *is identified and*

$$F_{Y^N,11}(y) = F_{Y,10}\big(F_{Y,00}^{-1}(F_{Y,01}(y))\big). \tag{9}$$

PROOF: By Assumption 3.2, $h(u, t)$ is invertible in u; denote this inverse by $h^{-1}(y; t)$. Consider the distribution $F_{Y^N,gt}$:

$$\begin{aligned} F_{Y^N,gt}(y) &= \Pr\big(h(U, t) \leq y | G = g, T = t\big) \\ &= \Pr\big(U \leq h^{-1}(y; t) | G = g, T = t\big) \\ &= \Pr(U \leq h^{-1}(y; t) | G = g) \\ &= \Pr(U_g \leq h^{-1}(y; t)) = F_{U,g}(h^{-1}(y; t)). \end{aligned} \tag{10}$$

The preceding equation is central to the proof and will be applied to all four combinations (g, t). First, letting $(g, t) = (0, 0)$ and substituting $y = h(u, 0)$,

$$F_{Y,00}(h(u, 0)) = F_{U,0}\big(h^{-1}(h(u, 0); 0)\big) = F_{U,0}(u).$$

Then applying $F_{Y,00}^{-1}$ to each side, we have, for all $u \in \mathbb{U}_0$,[16]

$$h(u, 0) = F_{Y,00}^{-1}(F_{U,0}(u)). \tag{11}$$

[16]Note that the support restriction is important here, because for $u \notin \mathbb{U}_0$, it is not necessarily true that $F_{Y,00}^{-1}(F_{Y,00}(h(u, 0))) = h(u, 0)$.

Second, applying (10) with $(g, t) = (0, 1)$, using the fact that $h^{-1}(y; 1) \in \mathbb{U}_0$ for all $y \in \mathbb{Y}_{01}$, and applying the transformation $F_{U,0}^{-1}(\cdot)$ to both sides,

$$F_{U,0}^{-1}(F_{Y,01}(y)) = h^{-1}(y; 1) \tag{12}$$

for all $y \in \mathbb{Y}_{01}$. Combining (11) and (12) yields, for all $y \in \mathbb{Y}_{01}$,

$$h(h^{-1}(y; 1), 0) = F_{Y,00}^{-1}(F_{Y,01}(y)). \tag{13}$$

Note that $h(h^{-1}(y; 1), 0)$ is the period 0 outcome for an individual with the realization of u that corresponds to outcome y in group 0 and period 1. Equation (13) shows that this outcome can be determined from the observable distributions.

Third, apply (10) with $(g, t) = (1, 0)$ and substitute $y = h(u, 0)$ to get

$$F_{Y,10}(h(u, 0)) = F_{U,1}(u). \tag{14}$$

Combining (13) and (14), and substituting into (10) with $(g, t) = (1, 1)$, we obtain, for all $y \in \mathbb{Y}_{01}$,

$$\begin{aligned} F_{Y^N,11}(y) &= F_{U,1}(h^{-1}(y; 1)) \\ &= F_{Y,10}\big(h(h^{-1}(y; 1), 0)\big) = F_{Y,10}\big(F_{Y,00}^{-1}(F_{Y,01}(y))\big). \end{aligned}$$

By Assumption 3.4 ($\mathbb{U}_1 \subseteq \mathbb{U}_0$), it follows that $\mathbb{Y}_{11}^N \subseteq \mathbb{Y}_{01}$. Thus, the directly estimable distributions $F_{Y,10}$, $F_{Y,00}$, and $F_{Y,01}$ determine $F_{Y^N,11}$ for all $y \in \mathbb{Y}_{11}^N$. *Q.E.D.*

Under the assumptions of the CIC model, we can interpret the identification result using a transformation

$$k^{\mathrm{CIC}}(y) = F_{Y,01}^{-1}(F_{Y,00}(y)). \tag{15}$$

This transformation gives the second-period outcome for an individual with an unobserved component u such that $h(u, 0) = y$. Then the distribution of Y_{11}^N is equal to the distribution of $k^{\mathrm{CIC}}(Y_{10})$. This transformation suggests that the average treatment effect can be written as

$$\begin{aligned} \tau^{\mathrm{CIC}} &\equiv \mathbb{E}[Y_{11}^I - Y_{11}^N] = \mathbb{E}[Y_{11}^I] - \mathbb{E}[k^{\mathrm{CIC}}(Y_{10})] \\ &= \mathbb{E}[Y_{11}^I] - \mathbb{E}\big[F_{Y,01}^{-1}(F_{Y,00}(Y_{10}))\big], \end{aligned} \tag{16}$$

and an estimator for this effect can be constructed using empirical distributions and sample averages.

The transformation k^{CIC} is illustrated in Figure 1. Start with a value of y, with associated quantile q in the distribution of Y_{10}, as illustrated in the bottom panel of Figure 1. Then find the quantile for the same value of y in the distribution of Y_{00}, $F_{Y,00}(y) = q'$. Next, compute the change in y according to k^{CIC} by

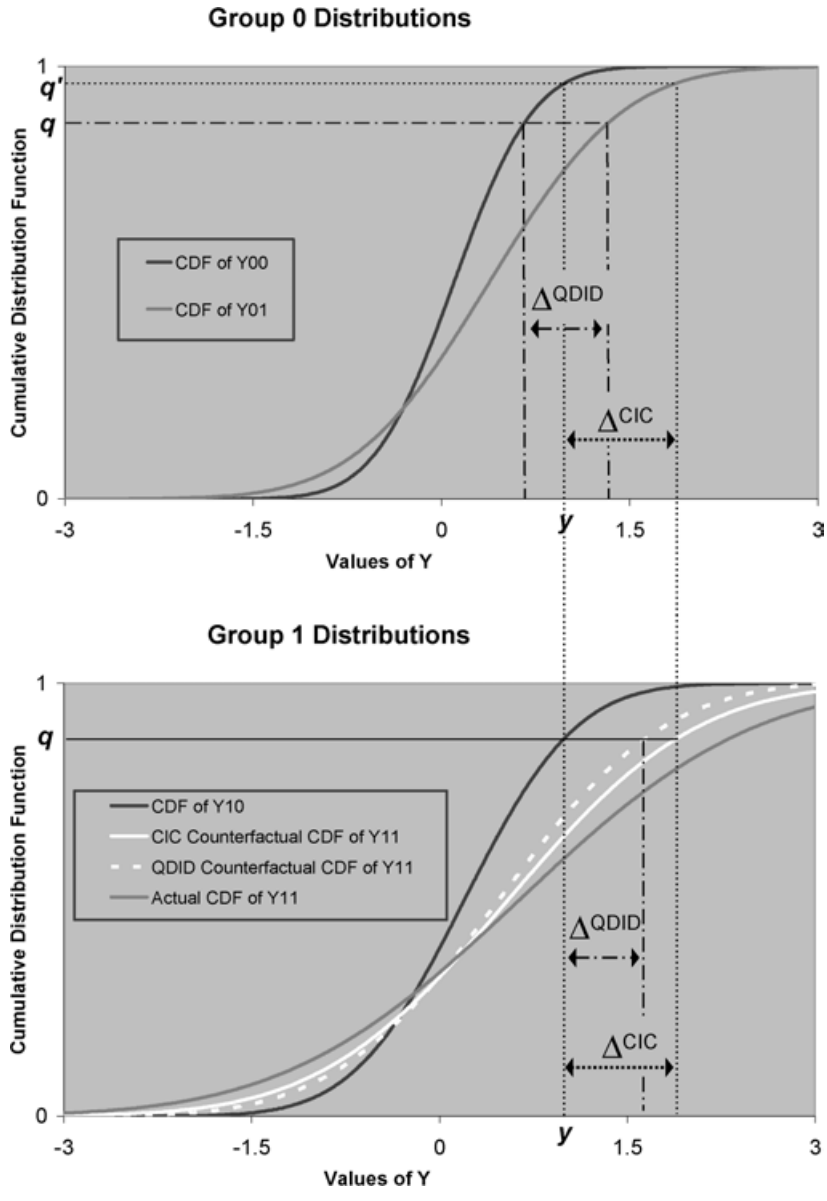

FIGURE 1.—Illustration of transformations.

finding the value for y at that quantile q' in the distribution of Y_{01} to get

$$\Delta^{\text{CIC}} = F^{-1}_{Y,01}(q') - F^{-1}_{Y,00}(q') = F^{-1}_{Y,01}(F_{Y,00}(y)) - y,$$

as illustrated in the top panel of Figure 1. Finally, compute a counterfactual

value of Y_{11}^N equal to $y + \Delta^{\text{CIC}}$, so that

$$k^{\text{CIC}}(y) = y + \Delta^{\text{CIC}} = F_{Y,01}^{-1}(F_{Y,00}(y)).$$

In contrast, for the standard DID model, the equivalent transformation is

$$k^{\text{DID}}(y) = y + \mathbb{E}[Y_{01}] - \mathbb{E}[Y_{00}].$$

Consider now the role of the support restriction, Assumption 3.4. Without it, we can only estimate the distribution function of Y_{11}^N on $\mathbb{Y}_{01}$. Outside that range, we have no information about the distribution of Y_{11}^N.

COROLLARY 3.1—Identification of the CIC Model Without Support Restrictions: *Suppose that Assumptions* 3.1–3.3 *hold, and that U is either continuous or discrete. Then we can identify the distribution of* Y_{11}^N *on* $\mathbb{Y}_{01}$. *For* $y \in \mathbb{Y}_{01}$, $F_{Y^N,11}$ *is given by* (9). *Outside of* $\mathbb{Y}_{01}$, *the distribution of* Y_{11}^N *is not identified.*

To see how this result could be used, suppose that Assumption 3.4 does not hold and $\mathbb{U}_1$ is not a subset of $\mathbb{U}_0$. Suppose also that $\mathbb{Y}_{00} = [\underline{y}_{00}, \overline{y}_{00}]$ so there are no holes in the support of Y_{00}. Define

$$\underline{q} = \min_{y \in \mathbb{Y}_{00}} F_{Y,10}(y), \quad \bar{q} = \max_{y \in \mathbb{Y}_{00}} F_{Y,10}(y). \tag{17}$$

Then, for any $q \in [\underline{q}, \bar{q}]$, we can calculate the effect of the treatment on quantile q of the distribution of $F_{Y,10}$ according to

$$\tau_q^{\text{CIC}} \equiv F_{Y^I,11}^{-1}(q) - F_{Y^N,11}^{-1}(q) = F_{Y^I,11}^{-1}(q) - F_{Y,01}^{-1}\big(F_{Y,00}(F_{Y,10}^{-1}(q))\big). \tag{18}$$

Thus, even without the support Assumption 3.4, for all quantiles of Y_{10} that lie in this range, it is possible to deduce the effect of the treatment. Furthermore, for any bounded function $g(y)$, it will be possible to put bounds on $\mathbb{E}[g(Y_{11}^I) - g(Y_{11}^N)]$, following the approach of Manski (1990, 1995). When g is the identity function and the supports are bounded, this approach yields bounds on the average treatment effect.

The standard DID approach requires no support assumption to identify the average treatment effect. Corollary 3.1 highlights the fact that the standard DID model identifies the average treatment effect only through extrapolation: because the average time trend is assumed to be the same in both groups, we can apply the time trend estimated on the control group to all individuals in the initial period treatment group, even those who experience outcomes outside the support of the initial period control group.

Also, observe that our analysis extends naturally to the case with covariates X; we simply require all assumptions to hold conditional on X. Then Theorem 3.1 extends to establish identification of $Y_{11}^N|X$ for realizations of X that are in the support of $X|G = g, T = t$ for each of $(g, t) \in \{\{0, 0\}, \{0, 1\}, \{1, 0\}\}$.

Of course, there is no requirement about how the distribution of X varies across subpopulations; thus, we can relax somewhat our assumption that population characteristics are stable over time within a group if all relevant factors that change over time are observable.

The CIC model treats groups and time periods asymmetrically. Of course, there is nothing intrinsic about the labels of period and group. In some applications, it might make more sense to reverse the roles of the two, yielding what we refer to as the reverse CIC (CIC-r) model. For example, CIC-r applies in a setting where, in each period, each member of a population is randomly assigned to one of two groups and these groups have different "production technologies." The production technology does not change over time in the absence of the intervention; however, the composition of the population changes over time (e.g., the underlying health of 60-year-old males participating in a medical study changes year by year), so that the distribution of U varies with time but not across groups. To uncover the average effect of the new technology, we need to estimate the counterfactual distribution in the second-period treatment group, which combines the treatment group production function with the second-period distribution of unobservables. When the distribution of outcomes is continuous, neither the CIC nor the CIC-r model has testable restrictions and so the two models cannot be distinguished. However, these approaches yield different estimates. Thus, it will be important in practice to justify the role of each dimension.

3.2. *The Counterfactual Effect of the Policy for the Untreated Group*

Until now, we have specified only a model for an individual's outcome in the absence of the intervention. No model for the outcome in the presence of the intervention is required to draw inferences about the effect of the policy change on the treatment group, that is, the effect of "the treatment on the treated" (e.g., Heckman and Robb (1985)); we simply need to compare the actual distribution of outcomes in the treated group with the counterfactual distribution inferred through the model for the outcomes in the absence of the treatment. However, more assumptions are required to analyze the effect of the treatment on the control group.

We augment the basic CIC model with an assumption about the treated outcomes. It seems natural to specify that these outcomes follow a model analogous to that for untreated outcomes, so that $Y^I = h^I(U, T)$. In words, at a given point in time, the effect of the treatment is the same across groups for individuals with the same value of the unobservable. However, outcomes can differ across individuals with different unobservables, and no further functional form assumptions are imposed on the incremental returns to treatment, $h^I(u, t) - h(u, t)$.[17]

[17]Although we require monotonicity of $h(u, t)$ and $h^I(u, t)$ in u, we do not require that the value of the unobserved component is identical in both regimes, merely that the distribution

At first, it might appear that finding the counterfactual distribution of Y_{01}^I could be qualitatively different than finding the counterfactual distribution of Y_{11}^N, because three out of four subpopulations did not experience the treatment. However, it turns out that the two problems are symmetric. Whereas $Y_{01}^I = h^I(U_0, 1)$ and $Y_{00} = h(U_0, 0)$,

$$Y_{01}^I \overset{d}{\sim} h^I(h^{-1}(Y_{00}; 0), 1). \tag{19}$$

To infer the distribution of Y_{01}^I it therefore suffices to represent the transformation $k(y) = h^I(h^{-1}(y; 0), 1)$ in terms of estimable functions. To do so, note that because the distribution of U_1 does not change with time, for $y \in \mathbb{Y}_{10}$,

$$F_{Y^I,11}^{-1}(F_{Y,10}(y)) = h^I(h^{-1}(y; 0), 1). \tag{20}$$

This is just the transformation $k^{\mathrm{CIC}}(y)$ with the roles of group 0 and group 1 reversed. Following this logic, to compute the counterfactual distribution of Y_{01}^I, we simply apply the approach outlined in Section 3.1, but replace G with $1 - G$.[18] Theorem 3.2 summarizes this concept:

THEOREM 3.2—Identification of the Counterfactual Effect of the Policy in the CIC Model: *Suppose that Assumptions* 3.1–3.3 *hold, and that U is either continuous or discrete. In addition, suppose that $Y^I = h^I(U, T)$, where $h^I(u, t)$ is strictly increasing in u. Then the distribution function of Y_{01}^I is identified on $\mathbb{Y}_{11}^I$ and is given by*

$$F_{Y^I,01}(y) = F_{Y,00}\big(F_{Y,10}^{-1}(F_{Y^I,11}(y))\big) \tag{21}$$

for all $y \in \mathbb{Y}_{11}^I$. If $\mathbb{U}_0 \subseteq \mathbb{U}_1$, then $\mathbb{Y}_{01}^I \subseteq \mathbb{Y}_{11}^I$, and $F_{Y^I,01}$ is identified everywhere.

PROOF: The proof is analogous to those of Theorem 3.1 and Corollary 3.1. Using (20), for $y \in \mathrm{supp}[Y_{11}^I]$,

$$F_{Y,10}^{-1}(F_{Y^I,11}(y)) = h(h^{I,-1}(y; 1), 0).$$

---

remains the same (that is, $U \perp T|G$). For example, letting U^N and U^I denote the unobserved components in the two regimes, we could have a fixed effect type error structure with $U_i^N = \varepsilon + \nu_i^N$ and $U_i^I = \varepsilon_i + \nu_i^I$, where the ε_i is a common component (fixed effect), and the ν_i^N and ν_i^I are idiosyncratic errors with the same distribution in both regimes.

[18]It might also be interesting to consider the effect that the treatment would have had in the first period. Our assumption that $h^I(u, t)$ can vary with t implies that Y_{00}^I and Y_{10}^I are not identified, because no information is available about $h^I(u, 0)$. Only if we make a much stronger assumption, such as $h^I(u, 0) = h^I(u, 1)$ for all u, can we identify the distribution of $Y_{g,0}^I$, but such an assumption would imply that $Y_{00}^I \overset{d}{\sim} Y_{01}^I$ and $Y_{10}^I \overset{d}{\sim} Y_{11}^I$, a fairly restrictive assumption. Comparably strong assumptions are required to infer the effect of the treatment on the control group in the CIC-r model, because the roles of group and time are reversed in that model.

Using this and (19), for $y \in \text{supp}[Y_{11}^I]$,

$$\begin{aligned}\Pr\big(h^I(h^{-1}(Y_{00};0),1) \le y\big) &= \Pr\big(Y_{00} \le F_{Y,10}^{-1}(F_{Y^I,11}(y))\big)\\ &= F_{Y,00}\big(F_{Y,10}^{-1}(F_{Y^I,11}(y))\big).\end{aligned}$$

The statement about supports follows from the definition of the model. *Q.E.D.*

Notice that in this model, not only can the policy change take place in a group with different distributional characteristics (e.g., "good" or "bad" groups tend to adopt the policy), but, furthermore, the expected benefit of the policy may vary across groups. Because $h^I(u,t) - h(u,t)$ varies with u, if $F_{U,0}$ is different from $F_{U,1}$, then the expected incremental benefit to the policy differs.[19] For example, suppose that $\mathbb{E}[h^I(U,1) - h(U,1)|G=1] > \mathbb{E}[h^I(U,1) - h(U,1)|G=0]$. Then, if the costs of adopting the policy are the same for each group, we would expect that if policies are chosen optimally, the policy would be more likely to be adopted in group 1. Using the method suggested by Theorem 3.2, it is possible to compare the average effect of the policy in group 1 with the counterfactual estimate of the effect of the policy in group 0 and to assess whether the group with the highest average benefits is indeed the one that adopted the policy. It is also possible to describe the range of adoption costs and distributions over unobservables for which the treatment would be cost-effective.

In the remainder of the paper, we focus on identification and estimation of the distribution of Y_{11}^N. However, the results that follow extend in a natural way to Y_{01}^I; simply exchange the labels of the groups 0 and 1 to calculate the negative of the treatment effect for group 0.

3.3. *The Quantile DID Model*

A third possible approach, distinct from the DID and CIC models, applies DID to each quantile rather than to the mean. We refer to this approach as the quantile DID approach (QDID). Some of the DID literature has followed this approach for specific quantiles, although it has not been studied as a method for obtaining the entire counterfactual distribution. For example, Poterba, Venti, and Wise (1995) and Meyer, Viscusi, and Durbin (1995) start from (1) and assume that the median of Y^N conditional on T and G is equal

[19]For example, suppose that the incremental returns to the intervention, $h^I(u,1) - h(u,1)$, are increasing in u, so that the policy is more effective for high-u individuals. If $F_{U,1}(u) \le F_{U,0}(u)$ for all u (i.e., first-order stochastic dominance), then expected returns to adopting the intervention are higher in group 1.

to $\alpha + \beta \cdot T + \gamma \cdot G$. Applying this approach to each quantile q, with the coefficients α_q, β_q, and γ_q indexed by the quantile, we obtain the transformation

$$k^{\text{QDID}}(y) = y + F^{-1}_{Y,01}(F_{Y,10}(y)) - F^{-1}_{Y,00}(F_{Y,10}(y))$$

with $F_{Y^N,11}(y) = \Pr(k^{\text{QDID}}(Y_{10}) \leq y)$. As illustrated in Figure 1, for a fixed y, we determine the quantile q for y in the distribution of Y_{10}, $q = F_{Y,10}(y)$. The difference over time in the control group at that quantile, $\Delta^{\text{QDID}} = F^{-1}_{Y,01}(q) - F^{-1}_{Y,00}(q)$, is added to y to get the counterfactual value, so that $k^{\text{QDID}}(y) = y + \Delta^{\text{QDID}}$. In this method, instead of comparing individuals across groups according to their outcomes and across time according to their quantiles, as in the CIC model, we compare individuals across both groups and time according to their quantile.

When outcomes are continuous, one can justify the QDID estimand using the model for the outcomes in the absence of the intervention,

$$Y^N = \tilde{h}(U, G, T) = \tilde{h}^G(U, G) + \tilde{h}^T(U, T), \tag{22}$$

in combination with the assumptions (i) $\tilde{h}(u, g, t)$ is strictly increasing in u and (ii) $U \perp (G, T)$. It is straightforward to see that the standard DID model is a special case of QDID.[20] Under the assumptions of the QDID model, the counterfactual distribution of Y^N_{11} is equal to that of $k^{\text{QDID}}(Y_{10})$. Details of the identification proof as well as an analysis of discrete-outcome case are in Athey and Imbens (2002) (hereafter AI).

Although the estimate of the counterfactual distribution under the QDID model differs from that under the DID model, under continuity the means of the two counterfactual distributions are identical: $\mathbb{E}[k^{\text{DID}}(Y_{10})] = \mathbb{E}[k^{\text{QDID}}(Y_{10})]$. The QDID model has several disadvantages relative to the CIC model: (i) additive separability of $\tilde{h}(u, g, t)$ is difficult to justify, in particular because it implies that the assumptions are not invariant to the scaling of y; (ii) the underlying distribution of unobservables must be identical in all subpopulations, eliminating an important potential source of intrinsic heterogeneity; (iii) the QDID model places some restrictions on the data.[21]

3.4. *Panel Data versus Repeated Cross Sections*

The discussion so far has avoided distinguishing between panel data and repeated cross sections. The presence of panel data creates some additional

[20] As with the CIC model, the assumptions of this model are unduly restrictive if outcomes are discrete. The discrete version of QDID allows $\tilde{h}(u, g, t)$ to be weakly increasing in u; the main substantive restriction implied by the QDID model is that the model should not predict outcomes out of bounds. For details on this case, see Athey and Imbens (2002).

[21] Without any restrictions on the distributions of Y_{00}, Y_{01}, and Y_{10}, the transformation k^{QDID} is not necessarily monotone, as it should be under the assumptions of the QDID model; thus, the model is testable (see AI for details).

possibilities. To discuss these issues it is convenient to modify the notation. For individual i, let Y_{it} be the outcome in period t for $t = 0, 1$. We allow the unobserved component U_{it} to vary with time:

$$Y_{it}^N = h(U_{it}, t).$$

The monotonicity assumption is the same as before: $h(u, t)$ must be increasing in u. We do not place any restrictions on the correlation between U_{i0} and U_{i1}, but we modify Assumption 3.3 to require that, conditional on G_i, the marginal distribution of U_{i0} is equal to the marginal distribution of U_{i1}. Formally, $U_{i0}|G_i \overset{d}{\sim} U_{i1}|G_i$. Note that the CIC model (like the standard DID model) does *not* require that individuals maintain their rank over time, that is, it does not require $U_{i0} = U_{i1}$. Although $U_{i0} = U_{i1}$ is an interesting special case, in many contexts, perfect correlation over time is not reasonable.[22] Alternatively, one may have a fixed effect specification $U_{it} = \varepsilon_i + \nu_{it}$, with ε_i a time-invariant individual-specific unobserved component (fixed effect) and ν_{it} an idiosyncratic error term with the same distribution in both periods.

The estimators proposed in this paper therefore apply to the panel setting as well as the cross-section setting. However, in panel settings there are additional methods available, including those developed for semiparametric models with fixed effects by Honore (1992), Kyriazidou (1997), and Altonji and Matzkin (1997, 2005). Another possibility in panel settings is to use the assumption of unconfoundedness or "selection on observables" (Rosenbaum and Rubin (1983), Barnow, Cain, and Goldberger (1980), Heckman and Robb (1985), Hirano, Imbens, and Ridder (2003)). Under such an assumption, individuals in the treatment group with an initial period outcome equal to y are matched to individuals in the control group with an identical first-period outcome, and their second-period outcomes are compared. Formally, let $F_{Y_{01}|Y_{00}}(\cdot|\cdot)$ be the conditional distribution function of Y_{01} given Y_{00}. Then, for the "selection on observables" model,

$$F_{Y^N,11}(y) = \mathbb{E}[F_{Y_{01}|Y_{00}}(y|Y_{10})],$$

which is in general different from the counterfactual distribution for the CIC model where $F_{Y^N,11}(y) = F_{Y,10}(F_{Y,00}^{-1}(F_{Y,01}(y)))$. The two models are equivalent if and only if $U_{i0} \equiv U_{i1}$, that is, if in the population there is perfect rank correlation between the first- and second-period unobserved components.

3.5. *Application to Wage Decompositions*

So far the focus has been on estimating the effect of interventions in settings with repeated cross sections and panels. A distinct but related problem arises

[22] If an individual gains experience or just age over time, her unobserved skill or health is likely to change.

in the literature on wage decompositions. In a typical example, researchers compare wage distributions for two groups, e.g., men and women or Whites and Blacks, at two points in time. Juhn, Murphy, and Pierce (1991) and Altonji and Blank (2000) decompose changes in Black–White wage differentials, after taking out differences in observed characteristics, into two effects: (i) the effect due to changes over time in the distribution of unobserved skills among Blacks and (ii) the effect due to common changes over time in the market price of unobserved skills.

In their survey of studies of race and gender in the labor market, Altonji and Blank (2000) formalize a suggestion by Juhn, Murphy, and Pierce (1991) to generalize the standard parametric, additive model for this problem to a nonparametric one, using the following assumptions: (i) the distribution of White skills does not change over time, whereas the distribution of Black skills can change in arbitrary ways; (ii) there is a single, strictly increasing function that maps skills to wages in each period—the market equilibrium pricing function. This pricing function can change over time, but is the same for both groups within a time period. Under the Altonji–Blank model, if we let Whites be group W and Blacks be group B, and let Y be the observed wage, then $\mathbb{E}[Y_{B1}] - \mathbb{E}[F^{-1}_{Y,W1}(F_{Y,W0}(Y_{B0}))]$ is interpreted as the part of the change in Blacks' average wages due to the change over time in unobserved Black skills. Interestingly, this expression is the same as the expression we derived for τ^{CIC}, even though the interpretation is different: in our case, the distribution of unobserved components remains the same over time and the difference is interpreted as the effect of an intervention.

Note that to apply an analog of our estimator of the effect of the treatment on the control group in the wage decomposition setting, we would require additional structure to specify what it would mean for Whites to experience "the same" change over time in their skill distribution that Blacks did, because the initial skill distributions are different. More generally, the precise relationship between estimands depends on the primitive assumptions for each model, because the CIC, CIC-r, and QDID models all lead to distinct estimands. The appropriateness of the assumptions of the underlying structural models must be justified in each application.

The asymptotic theory we provide for the CIC estimator can directly be applied to the wage decomposition problem as well. In addition, as we show below, the model, estimator, and asymptotic theory must be modified when data are discrete. Discrete wage data are common, because they arise if wages are measured in intervals or if there are mass points (such as the minimum wage, round numbers, or union wages) in the observed wage distribution.

3.6. *Relationship to Econometric Literature that Exploits Monotonicity*

In our approach to nonparametric identification, monotonicity of the production function plays a central role. Here, we build on Matzkin (1999, 2003),

who initiated a line of research that investigated the role of monotonicity in a wide range of models with an analysis of the case with exogenous regressors. In subsequent work (e.g., Das (2001, 2004), Imbens and Newey (2001), and Chesher (2003)), monotonicity of the relationship between the endogenous regressor and the unobserved component plays a crucial role in settings with endogenous regressors. In all these cases, as in the current paper, monotonicity in unobserved components implies a direct one-to-one link between the structural function and the distribution of the unobservables, a link that can be exploited in various ways. Most of these papers require strict monotonicity, typically ruling out settings with discrete endogenous regressors. An exception is Imbens and Angrist (1994), who used a weak monotonicity assumption and obtained results in the binary endogenous variable case for the subpopulation of compliers. One reason few results are available for binary or discrete data is that typically (as in this paper) discrete data in combination with weak monotonicity lead to loss of point identification of the usual estimands, e.g., population average effects. In the current paper, we show below that, although point identification is lost, one can still identify bounds on the population average effect of the intervention in the DID setting or regain point identification through additional assumptions.

Consider more specifically the relationship of our paper with the recent innovative work of Altonji and Matzkin (1997, 2005) (henceforth AM). In both our study and in AM, there is a central role for analyzing subpopulations that have the same distribution of unobservables. In our work, we argue that a defining feature of a group in a DID setting should be that the distribution of unobservables is the same in the group in different time periods. Altonji and Matzkin focus on subsets of the support of a vector of covariates Z, where, conditional on Z being in such a particular subset, the unobservables are independent of Z. In one example, Z incorporates an individual's history of experiences; permutations of that history should not affect the distribution of unobservables. So, an individual who completed first training program A and then program B would have the same unobservables as an individual who completed program B and then A. In a cross-sectional application, if in a given family, one sibling was a high-school graduate and the other a college graduate, both siblings would have the same unobservables. In both our study and in AM, within a subpopulation (induced by covariates) with a common distribution of unobservables, after normalizing the distribution of unobservables to be uniform, it is possible to identify a strictly increasing production function as the inverse of the distribution of outcomes conditional on the covariate. Altonji and Matzkin focus on estimation and inference for the production function itself, and for this they use an approach based on kernel methods. In contrast, we are interested in estimating the average difference of the production function for different subpopulations. We establish uniform convergence of our implicit estimator of the production function, so as to obtain root-n consistency of our estimator of the average treatment effect for the treated and control

groups as well as for treatment effects at a given quantile. We use the empirical distribution function, which does not require the choice of smoothing parameters, as an estimator of the distribution function of outcomes in each subpopulation. Furthermore, our approach generalizes naturally to the case with discrete outcomes (as we argue, a commonly encountered case) and our continuous-outcome estimator of the average treatment effect can be interpreted as a bound on the average treatment effect when outcomes are discrete. Thus, the researcher need not make an a priori choice about whether to use the discrete or the continuous model, because we provide bounds that collapse when outcomes are continuous.

4. IDENTIFICATION IN MODELS WITH DISCRETE OUTCOMES

In this section we consider the case with discrete outcomes. To simplify some of the subsequent arguments we assume that Y_{00} takes on only a finite number of values.

ASSUMPTION 4.1: *The random variable Y_{00} is discrete with a finite number of outcomes*: $\mathbb{Y}_{00} = \{\lambda_0, \ldots, \lambda_L\}$.

With discrete outcomes, the baseline CIC model as defined by Assumptions 3.1–3.3 is extremely restrictive. We therefore weaken Assumption 3.2 by allowing for weak rather than strict monotonicity. We show that this model is not point identified without additional assumptions and we calculate bounds on the counterfactual distribution. We also propose two approaches to tighten the bounds or even restore point identification: the first uses an additional assumption on the conditional distribution of unobservables and the second is based on the presence of exogenous covariates.[23]

4.1. *Bounds in the Discrete CIC Model*

The standard DID model implicitly imputes the average outcome in the second period for the treated subpopulation in the absence of the treatment with $\mathbb{E}[Y_{11}^N] = \mathbb{E}[Y_{10}] + (\mathbb{E}[Y_{01}] - \mathbb{E}[Y_{00}])$. With binary data, the imputed average for the second-period treatment group outcome is not guaranteed to lie in the interval $[0, 1]$. For example, suppose $\mathbb{E}[Y_{10}] = 0.5$, $\mathbb{E}[Y_{00}] = 0.8$, and $\mathbb{E}[Y_{01}] = 0.2$. In the control group, the probability of success decreases from 0.8 to 0.2. However, it is impossible that a similar percentage point decrease could

[23]However, there are other possible approaches for tightening the bounds. For example, one may wish to consider alternative restrictions on how the distribution of the unobserved components varies across groups, including stochastic dominance relationships or parametric functional forms. Alternatively, one may wish to put more structure on (the changes over time in) the production functions or restrict the treatment effect as a function of the unobserved component. We leave these possibilities for future work.

have occurred in the treated group in the absence of the treatment, because the implied probability of success would be less than zero.[24] The CIC model is also not very attractive, because it severely restricts the joint distribution of the observables.[25]

We therefore weaken the strict monotonicity condition as follows:

ASSUMPTION 4.2—Weak Monotonicity: *The function $h(u, t)$ is nondecreasing in u.*

We also assume continuity of U_0 and U_1:

ASSUMPTION 4.3—Continuity of U_0 and U_1: *The variables U_0 and U_1 are continuously distributed.*

The monotonicity assumption allows, for example, a latent index model $h(U, T) = \mathbb{1}\{\breve{h}(U, T) > 0\}$, for some $\breve{h}$ strictly increasing in U. With weak instead of strict monotonicity, we no longer obtain point identification. Instead, we can derive bounds on the average effect of the treatment in the spirit of Manski (1990, 1995). To build intuition, consider again an example with binary outcomes, $\mathbb{Y}_{gt} = \{0, 1\}$ for all g, t. Without loss of generality we assume $U_0 \sim \mathcal{U}[0, 1]$. Let $u^0(t) = \sup\{u : h(u, t) = 0\}$ so that

$$\mathbb{E}[Y_{gt}^N] = \Pr(U_g > u^0(t)). \tag{23}$$

In particular, $\mathbb{E}[Y_{11}^N] = \Pr(U_1 > u^0(1))$. All information regarding the distribution of U_1 is contained in the equality $\mathbb{E}[Y_{10}] = \Pr(U_1 > u^0(0))$. Suppose that $\mathbb{E}[Y_{01}] > \mathbb{E}[Y_{00}]$, implying $u^0(1) < u^0(0)$. Then there are two extreme cases for the conditional distribution of U_1 given $U_1 < u^0(0)$. First, all of the mass might be concentrated in the interval $[u^0(1), u^0(0)]$. In that case, $\Pr(U_1 > u^0(1)) = 1$. Second, there might be no mass between $u^0(1)$ and $u^0(0)$, in which case $\Pr(U_1 > u^0(1)) = \Pr(U_1 > u^0(0)) = \mathbb{E}[Y_{10}]$. Together, these two cases imply $\mathbb{E}[Y_{11}^N] \in [\mathbb{E}[Y_{10}], 1]$. Analogous arguments imply $\mathbb{E}[Y_{11}^N] \in [0, \mathbb{E}[Y_{10}]]$ when $\mathbb{E}[Y_{01}] < \mathbb{E}[Y_{00}]$. When $\mathbb{E}[Y_{01}] = \mathbb{E}[Y_{00}]$, we conclude that the production function does not change over time and neither does the probability of success

[24] One approach that has been used to deal with this problem (Blundell, Costa Dias, Meghir, and Van Reenen (2001)) is to specify an additive latent index model

$$Y_i = \mathbb{1}\{\alpha + \beta \cdot T_i + \eta \cdot G_i + \tau \cdot I_i + \varepsilon_i \geq 0\}.$$

Given a distributional assumption on ε_i (e.g., logistic), one can estimate the parameters of the latent index model and derive the implied estimated average effect for the second-period treatment group.

[25] For example, with binary outcomes, strict monotonicity of $h(u, t)$ in u implies that U is binary with $h(0, t) = 0$ and $h(1, t) = 1$, and thus $\Pr(Y = U|T = t) = 1$ or $\Pr(Y = U) = 1$. Independence of U and T then implies independence of Y and T, which is very restrictive.

change over time within a group, implying $\mathbb{E}[Y_{11}^N] = \mathbb{E}[Y_{10}]$. Whereas the average treatment effect is defined as $\tau = \mathbb{E}[Y_{11}^I] - \mathbb{E}[Y_{11}^N]$, it follows that

$$\tau \in \begin{cases} \big[\mathbb{E}[Y_{11}^I] - 1, \mathbb{E}[Y_{11}^I] - \mathbb{E}[Y_{10}]\big], & \text{if } \mathbb{E}[Y_{01}] > \mathbb{E}[Y_{00}], \\ \mathbb{E}[Y_{11}^I] - \mathbb{E}[Y_{10}], & \text{if } \mathbb{E}[Y_{01}] = \mathbb{E}[Y_{00}], \\ \big[\mathbb{E}[Y_{11}^I] - \mathbb{E}[Y_{10}], \mathbb{E}[Y_{11}^I]\big], & \text{if } \mathbb{E}[Y_{01}] < \mathbb{E}[Y_{00}]. \end{cases}$$

In this binary example the sign of the treatment effect is determined if and only if the observed time trends in the treatment and control groups move in opposite directions or if there is no time trend in the control group.

Now consider the general finite discrete case. Our definition of the inverse distribution function $F_Y^{-1}(q) = \inf\{y \in \mathbb{Y} | F_Y(y) \geq q\}$ implies $F_Y(F_Y^{-1}(q)) \geq q$. It is useful to have an alternative inverse distribution function. Define

$$F_Y^{(-1)}(q) = \sup\{y \in \mathbb{Y} \cup \{-\infty\} : F_Y(y) \leq q\}, \tag{24}$$

where we use the convention $F_Y(-\infty) = 0$. Define $\mathbb{Q} = \{q \in [0,1] | \exists\, y \in \mathbb{Y} \text{ s.t. } F_Y(y) = q\}$. For $q \in \mathbb{Q}$, the two definitions of inverse distribution functions agree so that $F_Y^{(-1)}(q) = F_Y^{-1}(q)$ and $F_Y^{-1}(F_Y(y)) = F_Y^{(-1)}(F_Y(y)) = y$. For $q \notin \mathbb{Q}$, $F_Y^{(-1)}(q) < F_Y^{-1}(q)$ and $F_Y(F_Y^{(-1)}(q)) < q$, so that, for all $q \in [0,1]$, we have $F_Y^{(-1)}(q) \leq F_Y^{-1}(q)$ and $F_Y(F_Y^{(-1)}(q)) \leq q \leq F_Y(F_Y^{-1}(q))$.

THEOREM 4.1—Bounds in the Discrete CIC Model: *Suppose that Assumptions* 3.1, 3.3, 3.4, 4.2, *and* 4.3 *hold*. *Then*

$$F_{Y^N,11}^{\mathrm{LB}}(y) \leq F_{Y^N,11}(y) \leq F_{Y^N,11}^{\mathrm{UB}}(y),$$

where, *for* $y < \inf \mathbb{Y}_{01}$, $F_{Y^N,11}^{\mathrm{LB}}(y) = F_{Y^N,11}^{\mathrm{UB}}(y) = 0$, *for* $y > \sup \mathbb{Y}_{01}$, $F_{Y^N,11}^{\mathrm{LB}}(y) = F_{Y^N,11}^{\mathrm{UB}}(y) = 1$, *and for* $y \in \mathbb{Y}_{01}$,

$$F_{Y^N,11}^{\mathrm{LB}}(y) = F_{Y,10}\big(F_{Y,00}^{(-1)}(F_{Y,01}(y))\big), F_{Y^N,11}^{\mathrm{UB}}(y) = F_{Y,10}\big(F_{Y,00}^{-1}(F_{Y,01}(y))\big). \tag{25}$$

These bounds are tight.

PROOF: By assumption, $\mathbb{U}_1 \subseteq \mathbb{U}_0$. Without loss of generality we can normalize U_0 to be uniform on $[0,1]$.[26] Then for $y \in \mathbb{Y}_{0t}$,

$$F_{Y,0t}(y) = \Pr(h(U_0, t) \leq y) = \sup\{u : h(u,t) = y\}.$$

[26]To see that there is no loss of generality, observe that, given that U is continuous, $F_{U,0}(u) = \Pr(F_{U,0}^{-1}(U_0^*) \leq u)$, where U_0^* is uniform on $[0,1]$. Then $\breve{h}(u,t) = h(F_{U,0}^{-1}(u), t)$ is nondecreasing in u because h is, and the distribution of Y_{0t} is unchanged. Whereas $\mathbb{U}_1 \subseteq \mathbb{U}_0$, the distribution of Y_{1t} is unchanged as well when we replace U_1 with $U_1^* \equiv F_{U_0}(U_1)$.

Using the normalization on U_0, we can express $F_{Y^N,11}(y)$ as

$$
\begin{aligned}
(26)\qquad F_{Y^N,1t}(y) &= \Pr(Y_{1t}^N \le y) = \Pr(h(U_1,t) \le y) \\
&= \Pr\bigl(U_1 \le \sup\{u : h(u,t) = y\}\bigr) = \Pr(U_1 \le F_{Y^N,0t}(y)).
\end{aligned}
$$

Using this and $F_Y(F^{(-1)}(q)) \le q \le F_Y(F_Y^{-1}(q))$,

$$
\begin{aligned}
(27)\qquad F_{Y,10}\bigl(F_{Y,00}^{(-1)}(F_{Y,01}(y))\bigr) &= \Pr\bigl(U_1 \le F_{Y,00}\bigl(F_{Y,00}^{(-1)}(F_{Y,01}(y))\bigr)\bigr) \\
&\le \Pr(U_1 \le F_{Y,01}(y)) = F_{Y^N,11}(y), \\
(28)\qquad F_{Y,10}\bigl(F_{Y,00}^{-1}(F_{Y,01}(y))\bigr) &= \Pr\bigl(U_1 \le F_{Y,00}\bigl(F_{Y,00}^{-1}(F_{Y,01}(y))\bigr)\bigr) \\
&\ge \Pr(U_1 \le F_{Y,01}(y)) = F_{Y^N,11}(y),
\end{aligned}
$$

which shows the validity of the bounds.

Next we show that the bounds are tight. We first construct a triple $(F_{U,0}(u), F_{U,1}^{\mathrm{LB}}(u), h(u,t))$ that is consistent with the distributions of Y_{00}, Y_{01}, and Y_{10}, and that leads to $F_{Y_{11}^N}^{\mathrm{LB}}(y)$ as the distribution function for Y_{11}^N. The choices are $U_0 \sim \mathcal{U}[0,1]$, $F_{U,1}^{\mathrm{LB}}(u) = F_{Y,10}(F_{Y,00}^{(-1)}(u))$, and $h(u,t) = F_{Y,0t}^{-1}(u)$. The choice is consistent with $F_{Y,0t}(y)$:

$$
\begin{aligned}
\Pr(Y_{0t} \le y) &= \Pr(h(U_0,t) \le y) = \Pr(F_{0t}^{-1}(U_0) \le y) \\
&= \Pr(U_0 \le F_{Y,0t}(y)) = F_{Y,0t}(y),
\end{aligned}
$$

where we rely on properties of inverse distribution functions stated in Lemma A.1 in the Appendix and proved in the supplement to this article. It is also consistent with $F_{Y,10}(y)$. First,

$$
\begin{aligned}
\Pr(Y_{10} \le y) &= \Pr(h(U_1,t) \le y) = \Pr(F_{Y,00}^{-1}(U_1) \le y) \\
&= \Pr(U_1 \le F_{Y,00}(y)) \\
&= F_{U,1}^{\mathrm{LB}}(F_{Y,00}(y)) = F_{Y,10}\bigl(F_{Y,00}^{(-1)}(F_{Y,00}(y))\bigr).
\end{aligned}
$$

At $y = \lambda_l \in \mathbb{Y}_{00}$ we have $F_{Y,00}^{(-1)}(F_{Y,00}(y)) = y$, so that $F_{Y,10}(F_{Y,00}^{(-1)}(F_{Y,00}(y))) = F_{Y,10}(y)$. If $\lambda_l < y < \lambda_{l+1}$, then $F_{Y,00}^{(-1)}(F_{Y,00}(y)) = \lambda_l$ and, because $\mathbb{Y}_{10} \subseteq \mathbb{Y}_{00}$, it follows that $F_{Y,10}(y) = F_{Y,10}(\lambda_l)$ so that again $F_{Y,10}(F_{Y,00}^{(-1)}(F_{Y,00}(y))) = F_{Y,10}(y)$. Finally, this choices leads to the distribution function for Y_{11}^N:

$$
\begin{aligned}
&F_{Y^N,11}(y) \\
&\quad = \Pr(h(U_1,1) \le y) = \Pr(F_{Y,01}^{-1}(U_1) \le y) = \Pr(U_1 \le F_{Y,01}(y)) \\
&\quad = F_{U,1}^{\mathrm{LB}}(F_{Y,01}(y)) = F_{Y,10}\bigl(F_{Y,00}^{(-1)}(F_{Y,01}(y))\bigr) = F_{Y^N,11}^{\mathrm{LB}}(y).
\end{aligned}
$$

This shows that $F_{Y^N,11}^{\mathrm{LB}}(y)$ is a tight lower bound on $F_{Y^N,11}(y)$

The argument that the upper bound is tight is more complicated. The difficulty is that we would like to choose the compound distribution function (c.d.f.) of U_1 to be $F^{\mathrm{UB}}_{U,1}(u) = F_{Y,10}(F^{-1}(u))$. However, this is not a distribution function in the discrete case, because it is not right continuous. However, we can approximate the upper bound $F^{\mathrm{UB}}_{Y^N,11}(y)$ arbitrarily closely by choosing $U_0 \sim \mathcal{U}[0,1]$, $h(u,t) = F^{-1}_{Y,0t}(u)$, and $F^{\mathrm{UB}}_{U,1}(u)$ close to $F_{Y,10}(F^{-1}(u))$. *Q.E.D.*

The proof of Theorem 4.1 is illustrated in Figure 2. The top left panel of the figure summarizes a hypothetical data set for an example with four possible outcomes, $\{\lambda_0, \lambda_1, \lambda_2, \lambda_3\}$. The top right panel of the figure illustrates the production function in each period, as inferred from the group 0 data (when U_0 is normalized to be uniform), where $u^k(t)$ is the value of u at which $h(u,t)$ jumps up to λ_k. In the bottom right panel, the diamonds represent the points of the distribution of U_1 that can be inferred from the distribution of Y_{10}. The distribution of U_1 is not identified elsewhere. This panel illustrates the infimum and supremum of the probability distributions that pass through the given points;

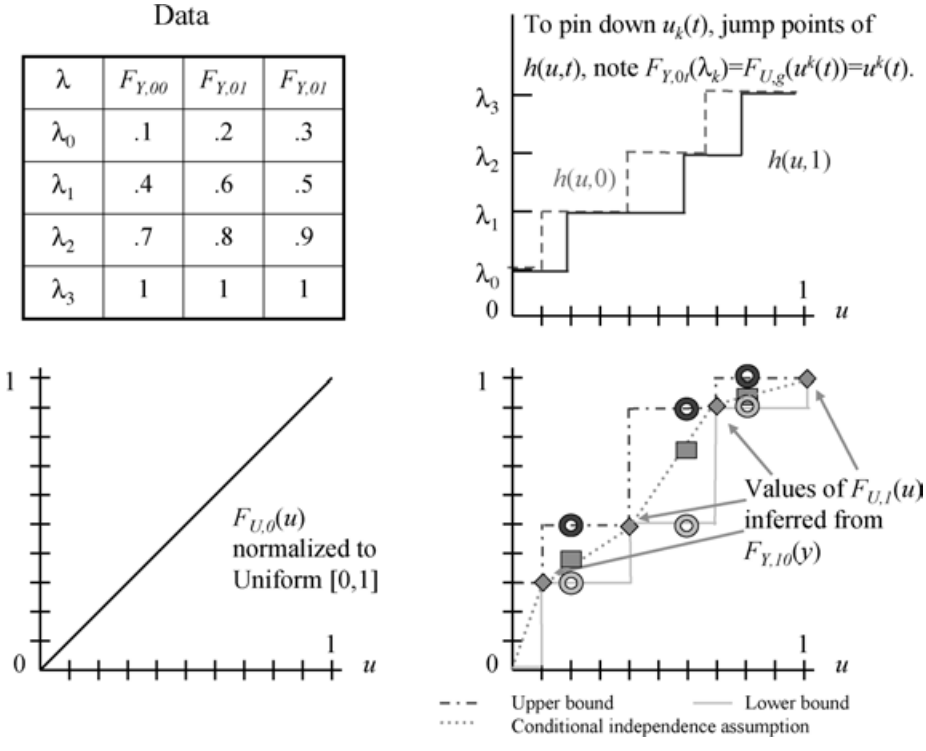

FIGURE 2.—Bounds and the conditional independence assumption in the discrete model.

these are bounds on F_{U_1}. The circles indicate the highest and lowest possible values of $F_{Y_{11}^N}(y) = F_{U_1}(u^k(t))$ for the support points; we will discuss the dotted line in the next section.

Note that if we simply ignore the fact that the outcome is discrete and use the continuous CIC estimator (9) to construct $F_{Y^N,11}$, we will obtain the upper bound $F^{\text{UB}}_{Y^N,11}$ from Theorem 4.1. If we calculate $\mathbb{E}[Y_{11}^N]$ directly from the distribution $F^{\text{UB}}_{Y^N,11}$,[27] we will thus obtain the *lower* bound for the estimate of $\mathbb{E}[Y_{11}^N]$, which in turn yields the *upper* bound for the average treatment effect, $\mathbb{E}[Y_{11}^I] - \mathbb{E}[Y_{11}^N]$.

The bounds are still valid under a weaker support condition. Instead of requiring that $\mathbb{U}_1 \subseteq \mathbb{U}_0$ (Assumption 3.4), it is sufficient that $\{\inf \mathbb{U}_1, \sup \mathbb{U}_1\} \subseteq \mathbb{U}_0$, which allows for the possibility of values in the support of the first-period treated distribution that are not in the support of the first-period control distribution, as long as these are not the boundary values.

4.2. *Point Identification in the Discrete CIC Model Through the Conditional Independence Assumption*

In combination with the previous assumptions, the following assumption restores point identification in the discrete CIC model.

ASSUMPTION 4.4—Conditional Independence: *We have* $U \perp G|Y, T$.

In the continuous CIC model, the level of outcomes can be compared across groups, and the quantile of outcomes can be compared over time. The role of Assumption 4.4 is to preserve that idea in the discrete model. In other words, to infer what would have happened to a treated unit in the first period with outcome y, we look at units in the first-period control group with the same outcome y. Using weak monotonicity, we can derive the distribution of their second-period outcomes (even if not their exact values as in the continuous case) and we use that to derive the counterfactual distribution for the second period treated in the absence of the intervention. Note that the strict monotonicity assumption (Assumption 3.2) implies Assumptions 4.2 and 4.4.[28]

To provide some intuition for the consequences of Assumption 4.4 for identification, we initially focus on the binary case. Without loss of generality normalize $U_0 \sim U[0, 1]$ and recall the definition of $u^0(t) = \sup\{u \in [0, 1] : h(u, t) = 0\}$,

[27] With continuous data, $k^{\text{CIC}}(Y_{10})$ has the distribution given in (9), and so (16) can be used to calculate the average treatment effect. As we show subsequently, with discrete data, $k^{\text{CIC}}(Y_{10})$ has distribution equal to $F^{\text{LB}}_{Y^N,11}$ rather than $F^{\text{UB}}_{Y^N,11}$, and so an estimate based directly on (9) yields a different answer than one based on (16).

[28] If $h(u, t)$ is strictly increasing in u, then one can write $U = h^{-1}(T, Y)$, so that, conditional on T and Y, the random variable U is degenerate and hence independent of G.

so that $1 - \mathbb{E}[Y_{gt}^N] = \Pr(U_g \leq u^0(t))$. Then we have, for $u \leq u^0(t)$,

$$\begin{aligned}\Pr(U_1 \leq u | U_1 \leq u^0(t)) &= \Pr\big(U_1 \leq u | U_1 \leq u^0(t), T = 0, Y = 0\big) \\ &= \Pr\big(U_0 \leq u | U_0 \leq u^0(t), T = 0, Y = 0\big) \\ &= \Pr(U_0 \leq u | U_0 \leq u^0(t)) = \frac{u}{u^0(t)}.\end{aligned}$$

Using the preceding expression together with an analogous expression for $\Pr(U_g > u |\ U_g > u^0(t))$ it is possible to derive the counterfactual $\mathbb{E}[Y_{11}^N]$:

$$\mathbb{E}[Y_{11}^N] = \begin{cases} \dfrac{\mathbb{E}[Y_{01}]}{\mathbb{E}[Y_{00}]}\mathbb{E}[Y_{10}] \\ \quad = \mathbb{E}[Y_{01}] + \dfrac{\mathbb{E}[Y_{01}]}{\mathbb{E}[Y_{00}]}(\mathbb{E}[Y_{10}] - \mathbb{E}[Y_{00}]) \\ \quad \text{if } \mathbb{E}[Y_{01}] \leq \mathbb{E}[Y_{00}], \\ 1 - \dfrac{1 - \mathbb{E}[Y_{01}]}{1 - \mathbb{E}[Y_{00}]}(1 - \mathbb{E}[Y_{10}]) \\ \quad = \mathbb{E}[Y_{01}] + \dfrac{1 - \mathbb{E}[Y_{01}]}{1 - \mathbb{E}[Y_{00}]}(\mathbb{E}[Y_{10}] - \mathbb{E}[Y_{00}]) \\ \quad \text{if } \mathbb{E}[Y_{01}] > \mathbb{E}[Y_{00}]. \end{cases}$$

Notice that this formula always yields a prediction for $\mathbb{E}[Y_{11}^N]$ between 0 and 1. When the time trend in the control group is negative, the counterfactual is the probability of successes in the treatment group initial period, adjusted by the proportional change over time in the probability of success in the control group. When the time trend is positive, the counterfactual probability of failure is the probability of failure in the treatment group in the initial period adjusted by the proportional change over time in the probability of failure in the control group.

The following theorem generalizes this discussion to more than two outcomes.

THEOREM 4.2—Identification of the Discrete CIC Model: *Suppose that Assumptions* 3.1, 3.3, 3.4, *and* 4.1–4.4 *hold. Suppose that the range of h is a discrete set* $\{\lambda_0, \ldots, \lambda_L\}$. *Then the distribution of* Y_{11}^N *is identified and is given by*

$$\begin{aligned}F_{Y^N,11}^{\mathrm{DCIC}}(y) = {} & F_{Y,10}\big(F_{Y,00}^{(-1)}(F_{Y,01}(y))\big) \\ & + \big(F_{Y,10}\big(F_{Y,00}^{-1}(F_{Y,01}(y))\big) - F_{Y,10}\big(F_{Y,00}^{(-1)}(F_{Y,01}(y))\big)\big) \\ & \times \frac{F_{Y,01}(y) - F_{Y,00}(F_{Y,00}^{(-1)}(F_{Y,01}(y)))}{F_{Y,00}(F_{Y,00}^{-1}(F_{Y,01}(y))) - F_{Y,00}(F_{Y,00}^{(-1)}(F_{Y,01}(y)))},\end{aligned} \tag{29}$$

if $F_{Y,00}(F_{Y,00}^{-1}(F_{Y,01}(y))) - F_{Y,00}(F_{Y,00}^{(-1)}(F_{Y,01}(y))) > 0$; *otherwise*, $F_{Y^N,11}^{\text{DCIC}}(y) = F_{Y,10}(F_{Y,00}^{(-1)}(F_{Y,01}(y)))$.

PROOF: We consider only the case with $F_{Y,00}(F_{Y,00}^{-1}(F_{Y,01}(y))) - F_{Y,00}(F_{Y,00}^{(-1)}(F_{Y,01}(y))) > 0$, because the other case is trivial. Without loss of generality we assume that $U_0 \sim \mathcal{U}[0,1]$. The proof exploits the fact that, for all $u \in [0,1]$ such that $u = F_{Y,00}(y)$ for some $y \in \mathbb{Y}_{00}$, we can directly infer the value of $F_{U,1}(u)$ as $F_{Y,10}(F_{Y,00}^{-1}(u))$ (or $F_{Y,10}(F_{Y,00}^{(-1)}(u))$, which is the same for such values of u). The first step is to decompose the distribution function of Y_{11}^N, using the fact that $F_{Y,00}(F_{Y,00}^{(-1)}(F_{Y,01}(y))) \le F_{Y,01}(y) \le F_{Y,00}(F_{Y,00}^{-1}(F_{Y,01}(y)))$:

$$
\begin{aligned}
F_{Y^N,11}(y) &= \Pr(Y_{11}^N \le y) = \Pr(h(U_1,1) \le y) = \Pr(U_1 \le F_{Y,01}(y)) \\
&= \Pr\big(U_1 \le F_{Y,00}\big(F_{Y,00}^{(-1)}(F_{Y,01}(y))\big)\big) \\
&\quad + \Pr\big(U_1 \le F_{Y,01}(y) \big| F_{Y,00}\big(F_{Y,00}^{(-1)}(F_{Y,01}(y))\big)\big) \\
&\le U_1 \le F_{Y,00}\big(F_{Y,00}^{-1}(F_{Y,01}(y))\big) \Pr\big(F_{Y,00}\big(F_{Y,00}^{(-1)}(F_{Y,01}(y))\big)\big) \\
&\le U_1 \le F_{Y,00}\big(F_{Y,00}^{-1}(F_{Y,01}(y))\big).
\end{aligned}
$$

Then we deal with the first term and the two factors in the second term separately. First,

$$
\Pr\big(U_1 \le F_{Y,00}\big(F_{Y,00}^{(-1)}(F_{Y,01}(y))\big)\big) = F_{Y,10}\big(F_{Y,00}^{(-1)}(F_{Y,01}(y))\big).
$$

Next,

$$
\begin{aligned}
&\Pr\big(F_{Y,00}\big(F_{Y,00}^{(-1)}(F_{Y,01}(y))\big) \le U_1 \le F_{Y,00}\big(F_{Y,00}^{-1}(F_{Y,01}(y))\big)\big) \\
&\quad = F_{Y,10}\big(F_{Y,00}^{-1}(F_{Y,01}(y))\big) - F_{Y,10}\big(F_{Y,00}^{(-1)}(F_{Y,01}(y))\big).
\end{aligned}
$$

Finally, using the conditional independence,

$$
\begin{aligned}
&\Pr\big(U_1 \le F_{Y,01}(y) \big| F_{Y,00}\big(F_{Y,00}^{(-1)}(F_{Y,01}(y))\big) \le U_1 \le F_{Y,00}\big(F_{Y,00}^{-1}(F_{Y,01}(y))\big)\big) \\
&\quad = \Pr\big(U_1 \le F_{Y,01}(y) | h(U_1,0) = F_{Y,00}^{-1}(F_{Y,01}(y))\big) \\
&\quad = \Pr\big(U_0 \le F_{Y,01}(y) | h(U_0,0) = F_{Y,00}^{-1}(F_{Y,01}(y))\big) \\
&\quad = \frac{F_{Y,01}(y) - F_{Y,0}(F_{Y,00}^{-1}(F_{Y,01}(y)))}{F_{Y,0}(F_{Y,00}^{-1}(F_{Y,01}(y))) - F_{Y,0}(F_{Y,00}^{(-1)}(F_{Y,01}(y)))}.
\end{aligned}
$$

Putting the three components together gives the desired result. *Q.E.D.*

The proof of Theorem 4.2 is illustrated in Figure 2. The dotted line in the bottom right panel illustrates the counterfactual distribution F_{U_1} based on the

conditional independence assumption. Given that U_0 is uniform, the conditional independence assumption requires the distribution of $U_1|Y = \lambda_l$ to be uniform for each l, and the point estimate of $F_{Y^N,11}(y)$ lies midway between the bounds of Theorem 4.1.

The average treatment effect, τ^{DCIC}, can be calculated using the distribution (29).

4.3. *Point Identification in the Discrete CIC Model Through Covariates*

In this subsection, we show that introducing observable covariates (X) can tighten the bounds on $F_{Y^N,11}$ and, with sufficient variation, can even restore point identification in the discrete-choice model without Assumption 4.4. The covariates are assumed to be independent of U conditional on the group, and the distribution of the covariates can vary with group and time.[29] Let $\mathbb{X}$ be the support of X, with $\mathbb{X}_{gt}$ the support of $X|G = g, T = t$. We assume that these supports are compact.

Let us modify the CIC model for the case of discrete outcomes with covariates.

ASSUMPTION 4.5—Discrete Model with Covariates: *The outcome of an individual in the absence of intervention satisfies the relationship*

$$Y^N = h(U, T, X).$$

ASSUMPTION 4.6—Weak Monotonicity: *The function $h(u, t, x)$ is nondecreasing in u and continuous in x for $t = 0, 1$ and for all $x \in \mathbb{X}$.*

ASSUMPTION 4.7—Covariate Independence: *We have $U \perp X|G$.*

We refer to the model defined by Assumptions 4.5–4.7, together with time invariance (Assumption 3.3), as the discrete CIC model with covariates. Note that Assumption 4.7 allows the distribution of X to vary with group and time.

To see how variation in X aids in identification, suppose that the range of h is the discrete set $\{\lambda_0, \ldots, \lambda_L\}$ and define

$$u^k(t, x) = \sup\{u' : h(u', t, x) \leq \lambda_k\}.$$

Recall that $F_{Y,10|X}(\cdot|x)$ reveals the value of $F_{U,1}(u)$ at all values $u \in \{u^0(t, x), \ldots, u^L(t, x)\}$, but nowhere else, as illustrated in Figure 2. Variation in X allows us to learn the value of $F_{U,1}(u)$ for more values of u.

[29]The assumption that $U \perp X|G$ is very strong. It should be carefully justified in applications using standards similar to those applied to justify instrumental variables. The analog of an "exclusion restriction" here is that X is excluded from $F_{U_g}(\cdot)$. Although the covariates can be time-varying, such variation can make the conditional independence of U even more restrictive.

More formally, define the functions $\underline{\mathcal{K}}: \mathbb{Y} \times \mathbb{X} \to \mathbb{Y}_{00} \cup \{-\infty\}$, $\underline{\mathcal{L}}: \mathbb{Y} \times \mathbb{X} \to \mathbb{X}_{00}$, $\overline{\mathcal{K}}: \mathbb{Y} \times \mathbb{X} \to \mathbb{Y}_{00}$, and $\overline{\mathcal{L}}: \mathbb{Y} \times \mathbb{X} \to \mathbb{X}_{00}$ by

$$\bigl(\underline{\mathcal{K}}(y;x), \underline{\mathcal{L}}(y;x)\bigr) = \arg \sup_{\substack{(y',x') \in (\mathbb{Y}_{00} \cup \{-\infty\}) \times \mathbb{X}_{00}: \\ F_{Y,00}(y'|x') \le F_{Y,01}(y|x)}} F_{Y,00}(y'|x'), \tag{30}$$

$$\bigl(\overline{\mathcal{K}}(y;x), \overline{\mathcal{L}}(y;x)\bigr) = \arg \inf_{\substack{(y',x') \in \mathbb{Y}_{00} \times \mathbb{X}_{00}: \\ F_{Y,00}(y'|x') \ge F_{Y,01}(y|x)}} F_{Y,00}(y'|x'). \tag{31}$$

If either of these is set-valued, take any element from the set of solutions. Because of the continuity in x and the finiteness of Y it follows that $F_{Y,00}(\underline{\mathcal{K}}(y;x)|\underline{\mathcal{L}}(y;x)) \le F_{Y,01}(y|x)$ and $F_{Y,00}(\overline{\mathcal{K}}(y;x)|\overline{\mathcal{L}}(y;x)) \ge F_{Y,01}(y|x)$.

The following result places bounds on the counterfactual distribution of Y_{11}^N.

THEOREM 4.3—Bounds in the Discrete CIC Model with Covariates: *Suppose that Assumptions* 3.3, 3.4, 4.3, *and* 4.5–4.7 *hold. Suppose that* $\mathbb{X}_{0t} = \mathbb{X}_{1t}$ *for* $t \in \{0, 1\}$. *Then we can place the following bounds on the distribution of* Y_{11}^N:

$$F^{\mathrm{LB}}_{Y^N,11|X}(y|x) = F_{Y|X,10}\bigl(\underline{\mathcal{K}}(y;x)|\underline{\mathcal{L}}(y;x)\bigr),$$

$$F^{\mathrm{UB}}_{Y^N,11|X}(y|x) = F_{Y|X,10}\bigl(\overline{\mathcal{K}}(y;x)|\overline{\mathcal{L}}(y;x)\bigr).$$

PROOF: Without loss of generality we normalize $U_0 \sim \mathcal{U}[0, 1]$. By continuity of U, we can express $F_{Y^N,1t}(y)$ as

$$\begin{aligned} F_{Y^N,1t|X}(y|x) &= \Pr(Y_{1t}^N \le y|X = x) = \Pr(h(U_1, t, x) \le y) \\ &= \Pr\bigl(U_1 \le \sup\{u : h(u, t, x) = y\}\bigr) \\ &= \Pr\bigl(U_1 \le F_{Y^N,0t|X}(y|x)\bigr). \end{aligned} \tag{32}$$

Thus, using (30) and (32),

$$\begin{aligned} F_{Y,10|X}\bigl(\underline{\mathcal{K}}(y;x)|\underline{\mathcal{L}}(y;x)\bigr) &= \Pr\bigl(U_1 \le F_{Y,00|X}\bigl(\underline{\mathcal{K}}(y;x)|\underline{\mathcal{L}}(y;x)\bigr)\bigr) \\ &\le \Pr\bigl(U_1 \le F_{Y,01|X}(y|x)\bigr) = F_{Y^N,11|X}(y|x), \\ F_{Y,10|X}\bigl(\overline{\mathcal{K}}(y;x)|\overline{\mathcal{L}}(y;x)\bigr) &= \Pr\bigl(U_1 \le F_{Y,00|X}\bigl(\overline{\mathcal{K}}(y;x)|\overline{\mathcal{L}}(y;x)\bigr)\bigr) \\ &\ge \Pr\bigl(U_1 \le F_{Y,01|X}(y|x)\bigr) = F_{Y^N,11}(y|x). \end{aligned}$$

Q.E.D.

When there is no variation in X, the bounds are equivalent to those given in Theorem 4.1. When there is sufficient variation in X, the bounds collapse and point identification can be restored.

THEOREM 4.4—Point Identification of the Discrete CIC Model with Covariates: *Suppose that Assumptions* 3.3, 3.4, 4.3, *and* 4.5–4.7 *hold*. *Suppose that* $\mathbb{X}_{0t} = \mathbb{X}_{1t}$ *for* $t \in \{0, 1\}$. *Define*

$$S_t(y) = \{u : \exists\, x \in \mathbb{X}_{0t} \text{ s.t. } u = F_{Y,0t|X}(y|x)\}. \tag{33}$$

Assume that, for all $y \in \mathbb{Y}_{01}$, $S_1(y) \subseteq \bigcup_{y \in \mathbb{Y}_{00}} S_0(y)$. *Then the distribution of* $Y_{11}^N|X$ *is identified*.

PROOF: Normalize $U_0 \sim \mathcal{U}[0, 1]$. For each $x \in \mathbb{X}_{01}$ and each $y \in \mathbb{Y}_{01}$, let $(\psi(y; x), \chi(y; x))$ be an element of the set of pairs $(y', x') \in \{\mathbb{Y}_{00}, \mathbb{X}_{00}\}$ that satisfy $F_{Y,00|X}(y'|x') = F_{Y,01|X}(y|x)$. Whereas $S_1(y) \subseteq \bigcup_{y \in \mathbb{Y}_{00}} S_0(y)$, there exist such a y' and x'. Then

$$\begin{aligned} F_{Y^N|X,11}(y|x) &= F_{U,1}(F_{Y,01|X}(y|x)) = F_{U,1}\big(F_{Y,00|X}(\psi(y; x)|\chi(y; x))\big) \\ &= F_{Y|X,10}(\psi(y; x)|\chi(y; x)). \end{aligned}$$

Q.E.D.

5. INFERENCE

In this section we consider inference for the continuous and discrete CIC models.

5.1. *Inference in the Continuous CIC Model*

To guarantee that $\tau^{\text{CIC}} = \mathbb{E}[Y_{11}^I] - \mathbb{E}[Y_{11}^N]$ is equal to $\mathbb{E}[Y_{11}] - \mathbb{E}[F_{Y,01}^{-1}(F_{Y,00}(Y_{10}))]$, we maintain Assumptions 3.1–3.4 in this subsection. Alternatively, we could simply redefine the parameter of interest as $\mathbb{E}[Y_{11}] - \mathbb{E}[F_{Y,01}^{-1}(F_{Y,00}(Y_{10}))]$, because those assumptions are not directly used in the analysis of inference. We make the following assumptions regarding the sampling process.

ASSUMPTION 5.1—Data Generating Process:

(i) *Conditional on* $T_i = t$ *and* $G_i = g$, Y_i *is a random draw from the subpopulation with* $G_i = g$ *during period* t.

(ii) *For all* $t, g \in \{0, 1\}$, $\alpha_{gt} \equiv \Pr(T_i = t, G_i = g) > 0$.

(iii) *The four random variables* Y_{gt} *are continuous with densities* $f_{Y,gt}(y)$ *that are continuously differentiable, bounded from above by* $\overline{f}_{gt}$, *and bounded from below by* $\underline{f}_{gt} > 0$ *with support* $\mathbb{Y}_{gt} = [\underline{y}_{gt}, \overline{y}_{gt}]$.

(iv) *We have* $\mathbb{Y}_{10} \subseteq \mathbb{Y}_{00}$.

We have four random samples, one from each group–period. Let the observations from group g and time period t be denoted by $Y_{gt,i}$ for $i = 1, \ldots, N_{gt}$. We use the empirical distribution as an estimator for the distribution function:

$$\hat{F}_{Y,gt}(y) = \frac{1}{N_{gt}} \sum_{i=1}^{N_{gt}} \mathbb{1}\{Y_{gt,i} \leq y\}. \tag{34}$$

As an estimator for the inverse of the distribution function, we use

$$\hat{F}^{-1}_{Y,gt}(q) = \inf\{y \in \mathbb{Y}_{gt} : \hat{F}_{Y,gt}(y) \geq q\}, \tag{35}$$

so that $F^{-1}_{Y,gt}(0) = \underline{y}_{gt}$. As an estimator of $\tau^{\text{CIC}} = \mathbb{E}[Y_{11}] - \mathbb{E}[F^{-1}_{Y,01}(F_{Y,00}(Y_{10}))]$, we use

$$\hat{\tau}^{\text{CIC}} = \frac{1}{N_{11}} \sum_{i=1}^{N_{11}} Y_{11,i} - \frac{1}{N_{10}} \sum_{i=1}^{N_{10}} \hat{F}^{-1}_{Y,01}(\hat{F}_{Y,00}(Y_{10,i})). \tag{36}$$

To present results on the large sample approximations to the sampling distribution of this estimator, we need a couple of additional definitions. First, define

$$P(y, z) = \frac{1}{f_{Y,01}(F^{-1}_{Y,01}(F_{Y,00}(z)))} \cdot (\mathbb{1}\{y \leq z\} - F_{Y,00}(z)), \tag{37}$$

$$p(y) = \mathbb{E}[P(y, Y_{10})],$$

$$Q(y, z) = -\frac{1}{f_{Y,01}(F^{-1}_{Y,01}(F_{Y,00}(z)))} \times \big(\mathbb{1}\{F_{Y,01}(y) \leq F_{Y,00}(z)\} - F_{Y,00}(z)\big), \tag{38}$$

$$q(y) = \mathbb{E}[Q(y, Y_{10})],$$

$$r(y) = F^{-1}_{Y,01}(F_{Y,00}(y)) - \mathbb{E}\big[F^{-1}_{Y,01}(F_{Y,00}(Y_{10}))\big], \tag{39}$$

$$s(y) = y - \mathbb{E}[Y_{11}], \tag{40}$$

with corresponding variances $V^p = \mathbb{E}[p(Y_{00})^2]$, $V^q = \mathbb{E}[q(Y_{01})^2]$, $V^r = \mathbb{E}[r(Y_{10})^2]$, and $V^s = \mathbb{E}[s(Y_{11})^2]$, respectively.

THEOREM 5.1—Consistency and Asymptotic Normality: *Suppose Assumption* 5.1 *holds. Then* (i) $\hat{\tau}^{\text{CIC}} - \tau^{\text{CIC}} = O_p(N^{-1/2})$ *and* (ii) $\sqrt{N}(\hat{\tau}^{\text{CIC}} - \tau^{\text{CIC}}) \xrightarrow{d} \mathcal{N}(0, V^p/\alpha_{00} + V^q/\alpha_{01} + V^r/\alpha_{10} + V^s/\alpha_{11})$.

See Appendix A for the proof.

An initial step in the proof is to linearize the estimator by showing that

$$\hat{\tau} = \tau + \frac{1}{N_{00}} \sum_{i=1}^{N_{00}} p(Y_{00,i}) + \frac{1}{N_{01}} \sum_{i=1}^{N_{01}} q(Y_{01,i}) + \frac{1}{N_{10}} \sum_{i=1}^{N_{10}} r(Y_{10,i}) + \frac{1}{N_{11}} \sum_{i=1}^{N_{11}} s(Y_{11,i}) + o_p(N^{-1/2}).$$

The variance of the CIC estimator can be equal to the variance of the standard DID estimator $\hat{\tau}^{\text{DID}} = \overline{Y}_{11} - \overline{Y}_{10} - (\overline{Y}_{01} - \overline{Y}_{00})$ in some special cases, such as when the following conditions hold: (i) Assumption 5.1, (ii) $Y_{00} \overset{d}{\sim} Y_{10}$, and (iii) for some $a \in \mathbb{R}$ and for $g = 0, 1$, $Y_{g0}^N \overset{d}{\sim} Y_{g1}^N + a$. More generally, the variance of $\hat{\tau}^{\text{CIC}}$ can be larger or smaller than the variance of $\hat{\tau}^{\text{DID}}$.[30]

To estimate the asymptotic variance $V^P/\alpha_{00} + V^q/\alpha_{01} + V^r/\alpha_{10} + V^s/\alpha_{11}$, we replace expectations with sample averages, using empirical distribution functions and their inverses for distribution functions and their inverses, and using any uniformly consistent nonparametric estimator for the density functions.[31] Specifically, given estimators for the conditional densities, we first estimate $P(y, z)$, $Q(y, z)$, $r(y)$, and $s(y)$ by substituting these estimators for $f_{Y,gt}(y)$, $F_{Y,gt}(y)$, and $F_{Y,gt}^{-1}(q)$, and sample averages for expectations. We then estimate $p(y)$ and $q(y)$ by $\hat{p}(y) = \sum_{i=1}^{N_{10}} \hat{P}(y, Y_{10,i})/N_{10}$ and $\hat{q}(y) = \sum_{i=1}^{N_{10}} \hat{Q}(y, Y_{10,i})/N_{10}$, respectively. Finally, we estimate V^p, V^q, V^r,

[30]To see this, suppose that Y_{00} has mean zero, unit variance, and compact support, and that $Y_{00} \overset{d}{\sim} Y_{10}$. Now suppose that $Y_{g1}^N \overset{d}{\sim} \sigma \cdot Y_{g0}$ for some $\sigma > 0$, and thus Y_{g1}^N has mean zero and variance σ^2 for each g. The assumptions of the both the CIC model and the mean-independence DID model are satisfied, and the probability limits of $\hat{\tau}^{\text{DID}}$ and $\hat{\tau}^{\text{CIC}}$ are identical and equal to $\mathbb{E}[Y_{11}] - \mathbb{E}[Y_{10}] - [\mathbb{E}[Y_{01}] - \mathbb{E}[Y_{00}]]$. If N_{00} and N_{01} are much larger than N_{10} and N_{11}, the variance of the standard DID estimator is essentially equal to $\text{Var}(Y_{11}) + \text{Var}(Y_{10})$. The variance of the CIC estimator is in this case approximately equal to $\text{Var}(Y_{11}) + \text{Var}(k(Y_{10})) = \text{Var}(Y_{11}) + \sigma^2 \cdot \text{Var}(Y_{10})$. Hence with $\sigma^2 < 1$, the CIC estimator is more efficient, and with $\sigma^2 > 1$, the standard DID estimator is more efficient.

[31]For example, to ensure that the estimator is uniformly consistent, including at the boundary points, let $\tilde{Y}_{gt}$ be the midpoint of the support, $\tilde{Y}_{gt} = (\overline{y}_{gt} - \underline{y}_{gt})/2$. Then we can use the estimator for $f_{Y,gt}(y)$:

$$\hat{f}_{Y,gt} = \begin{cases} (\hat{F}_{Y,gt}(y + N^{-1/3}) - \hat{F}_{Y,gt}(y))/N^{-1/3}, & \text{if } y \le \tilde{Y}_{gt}, \\ (\hat{F}_{Y,gt}(y) - \hat{F}_{Y,gt}(y - N^{-1/3}))/N^{-1/3}, & \text{if } y > \tilde{Y}_{gt}. \end{cases}$$

Other estimators for $\hat{f}_{Y,gt}(y)$ can be used as long as they are uniformly consistent, including at the boundary of the support.

and V_s as

$$\hat{V}^p = \frac{1}{N_{00}} \sum_{i=1}^{N_{00}} \hat{p}(Y_{00,i})^2, \quad \hat{V}^q = \frac{1}{N_{01}} \sum_{i=1}^{N_{01}} \hat{q}(Y_{01,i})^2,$$

$$\hat{V}^r = \frac{1}{N_{10}} \sum_{i=1}^{N_{10}} \hat{r}(Y_{10,i})^2, \quad \hat{V}^s = \frac{1}{N_{11}} \sum_{i=1}^{N_{11}} \hat{s}(Y_{11,i})^2,$$

and estimate α_{gt} by $\hat{\alpha}_{gt} = \sum_i \mathbb{1}\{G_i = g, T_i = t\}/N$.

THEOREM 5.2—Consistent Estimation of the Variance: *Suppose Assumption* 5.1 *holds. Then* $\hat{\alpha}_{gt} \xrightarrow{p} \alpha_{gt}$ *for all* g, t, $\hat{V}^p \xrightarrow{p} V^p$, $\hat{V}^q \xrightarrow{p} V^q$, $\hat{V}^r \xrightarrow{p} V^r$, $\hat{V}^s \xrightarrow{p} V^s$, *and, therefore,*

$$\hat{V}^p/\hat{\alpha}_{00} + \hat{V}^q/\hat{\alpha}_{01} + \hat{V}^r/\hat{\alpha}_{10} + \hat{V}^s/\hat{\alpha}_{11} \xrightarrow{p} V^p/\alpha_{00} + V^q/\alpha_{01} + V^r/\alpha_{10} + V^s/\alpha_{11}.$$

See Appendix A for the proof.

For the quantile case we estimate τ_q^{CIC} as

$$\hat{\tau}_q^{\text{CIC}} = \hat{F}_{Y,11}^{-1}(q) - \hat{F}_{Y,01}^{-1}\big(\hat{F}_{Y,00}(\hat{F}_{Y,10}^{-1}(q))\big).$$

To establish its asymptotic properties, it is useful to define the quantile analog of the functions $p(\cdot)$, $q(\cdot)$, $r(\cdot)$, and $s(\cdot)$, denoted by $p_q(\cdot)$, $q_q(\cdot)$, $r_q(\cdot)$, and $s_q(\cdot)$:

$$\begin{aligned}
p_q(y) &= \frac{1}{f_{Y,01}(F_{Y,01}^{-1}(F_{Y,00}(F_{Y,10}^{-1}(q))))} \\
&\quad \times \big(\mathbb{1}\{y \le F_{Y,10}^{-1}(q)\} - F_{Y,00}(F_{Y,10}^{-1}(q))\big), \\
q_q(y) &= -\frac{1}{f_{Y,01}(F_{Y,01}^{-1}(F_{Y,00}(F_{Y,10}^{-1}(q))))} \\
&\quad \times \big(\mathbb{1}\big\{F_{Y,01}(y) \le F_{Y,00}(F_{Y,10}^{-1}(q))\big\} - F_{Y,00}(F_{Y,10}^{-1}(q))\big), \\
r_q(y) &= -\frac{f_{Y,00}(F_{Y,10}^{-1}(q))}{f_{Y,01}(F_{Y,01}^{-1}(F_{Y,00}(F_{Y,10}^{-1}(q))))f_{Y,10}(F_{Y,10}^{-1}(q))} \\
&\quad \times \big(\mathbb{1}\{F_{Y,10}(y) \le q\} - q\big), \\
s_q(y) &= -\frac{1}{f_{Y,11}(F_{Y,11}^{-1}(q))}\big(\mathbb{1}\{y \le F_{Y,11}^{-1}(q)\} - q\big),
\end{aligned}$$

with corresponding variances $V_q^p = \mathbb{E}[p_q(Y_{00})^2]$, $V_q^q = \mathbb{E}[q_q(Y_{01})^2]$, $V_q^r = \mathbb{E}[r_q(Y_{10})^2]$, and $V_q^s = \mathbb{E}[s_q(Y_{11})^2]$.

THEOREM 5.3—Consistency and Asymptotic Normality of Quantile CIC Estimator: *Suppose Assumption* 5.1(i)–(iii) *hold. Then, defining* $\underline{q}$ *and* $\bar{q}$ *as in* (17), *for all* $q \in (\underline{q}, \bar{q})$:

(i) $\hat{\tau}_q^{\text{CIC}} \xrightarrow{p} \tau_q^{\text{CIC}}$,

(ii) $\sqrt{N}(\hat{\tau}_q^{\text{CIC}} - \tau_q^{\text{CIC}}) \xrightarrow{d} \mathcal{N}(0, V_q^p/\alpha_{00} + V_q^q/\alpha_{01} + V_q^r/\alpha_{10} + V_q^s/\alpha_{11})$.

See the supplement (Athey and Imbens (2006)) for the proof.

The variance of the quantile estimators can be estimated analogously to that for the estimator of the average treatment effect.

We may also wish to test the null hypothesis that the treatment has no effect by comparing the distributions of the second-period outcome for the treatment group with and without the treatment—that is, $F_{Y^I,11}(y)$ and $F_{Y^N,11}(y)$—or test for first- or second-order stochastic dominance relationships (e.g., Abadie (2002)). One approach for testing the equality hypothesis is to estimate $\hat{\tau}_q^{\text{CIC}}$ for a number of quantiles and jointly test their equality. For example, one may wish to estimate the three quartiles or the nine deciles and test whether they are identical in the distributions of Y_{11}^I and Y_{11}^N. In AI, we provide details about carrying out such a test, showing that a $\mathcal{X}^2$ test can be used. More generally, it may be possible to construct a Kolmogorov–Smirnov or Cramer–Von Mises test on the entire distribution. Such tests could be used to test the assumptions that underlie the model if more than two time periods are available.

With discrete covariates, one can estimate the average treatment effect for each value of the covariates by applying the estimator discussed in Theorem 5.1 and taking the average over the distribution of the covariates. When the covariates take on many values, this procedure may be infeasible and one may wish to smooth over different values of the covariates. One approach is to estimate the distribution of each Y_{gt} nonparametrically conditional on covariates X (using kernel regression or series estimation) and then again average the average treatment effect at each X over the appropriate distribution of the covariates.

As an alternative, consider a more parametric approach to adjusting for covariates. Suppose

$$h(u, t, x) = h(u, t) + x'\beta \quad \text{and} \quad h^I(u, t, x) = h^I(u, t) + x'\beta$$

with U independent of (T, X) given G.[32] In this model the effect of the in-

[32]A natural extension would consider a model of the form $h(u, t) + m(x)$; the function m could be estimated using nonparametric regression techniques, such as series expansion or kernel regression. Alternatively, one could allow the coefficients β to depend on the group and/or time. The latter extension would be straightforward given the results in AI.

tervention does not vary with X (although it still varies by unobserved differences between units). The average treatment effect is given by $\tau^{\text{CIC}} = \mathbb{E}[\tilde{Y}_{11}] - \mathbb{E}[F^{-1}_{\tilde{Y},01}(F_{\tilde{Y},00}(\tilde{Y}_{10}))]$, where $\tilde{Y}_{gt,i} = Y_{gt,i} - X'_{gt,i}\beta$. To derive an estimator for τ^{CIC}, we proceed as follows. First, β can be estimated consistently using linear regression of outcomes on X and the four group–time dummy variables (without an intercept). We can then apply the CIC estimator to the residuals from an ordinary least squares regression with the effects of the dummy variables added back in. To be precise, define $D = ((1-T)(1-G), T(1-G), (1-T)G, TG)'$. In the first stage, we estimate the regression

$$Y_i = D'_i\delta + X'_i\beta + \varepsilon_i.$$

Then construct the residuals with the group–time effects left in:

$$\hat{Y}_i = Y_i - X'_i\hat{\beta} = D'_i\hat{\delta} + \hat{\varepsilon}_i.$$

Finally, apply the CIC estimator to the empirical distributions of the augmented residuals $\hat{Y}_i$. In AI we show that this covariance-adjusted estimator of τ^{CIC} is consistent and asymptotically normal, and we calculate the asymptotic variance.

5.2. *Inference in the Discrete CIC Model*

In this subsection we discuss inference for the discrete CIC model. If one is willing to make the conditional independence assumption, Assumption 4.4, the model is a fully parametric model and inference becomes standard using likelihood methods. We therefore focus on the discrete case without Assumption 4.4. We maintain Assumptions 3.1, 3.3, 3.4, and 4.2 (as in the continuous case, these assumptions are used only for the interpretation of the bounds τ_{LB} and τ_{UB}, and they are not used directly in the analysis of inference). We make one additional assumption.

ASSUMPTION 5.2—Absence of Ties: *We have that* $\mathbb{Y}$ *is a finite set and, for all* $y, y' \in \mathbb{Y}$,

$$F_{Y,01}(y) \neq F_{Y,00}(y').$$

If, for example, $\mathbb{Y} = \{0, 1\}$, this assumption requires $\Pr(Y_{01} = 0) \neq \Pr(Y_{00} = 0)$ and $\Pr(Y_{01} = 0), \Pr(Y_{00} = 0) \in (0, 1)$. When ties of this sort are not ruled out, the bounds on the distribution function do not converge to their theoretical values as the sample size increases.[33]

[33] An analogous situation arises in estimating the median of a binary random variable Z with $\Pr(Z = 1) = p$. If $p \neq 1/2$, the sample median will converge to the true median (equal to $\mathbb{1}\{p \geq 1/2\}$), but if $p = 1/2$, then in large samples the estimated median will be equal to 1 with probability $1/2$ and equal to 0 with probability $1/2$.

Define

$$\underline{F}_{Y,00}(y) = \Pr(Y_{00} < y), \tag{41}$$

$$\underline{k}(y) = F^{-1}_{Y,01}(\underline{F}_{Y,00}(y)), \quad \text{and} \quad \overline{k}(y) = F^{-1}_{Y,01}(F_{Y,00}(y)) \tag{42}$$

with estimated counterparts

$$\hat{\underline{F}}_{Y,00}(y) = \frac{1}{N_{00}} \sum_{i=1}^{N_{00}} \mathbb{1}\{Y_{00,i} < y\}, \tag{43}$$

$$\hat{\underline{k}}(y) = \hat{F}^{-1}_{Y,01}(\hat{\underline{F}}_{Y,00}(y)), \quad \text{and} \quad \hat{\overline{k}}(y) = \hat{F}^{-1}_{Y,01}(\hat{F}_{Y,00}(y)). \tag{44}$$

The functions $\underline{k}(y)$ and $\overline{k}(y)$ can be interpreted as the bounds on the transformation $k(y)$ defined for the continuous case in (15). Note that $\overline{k}(y) \equiv k^{\text{CIC}}(y)$. In the Appendix (Lemma A.12), we show that the c.d.f. of $\underline{k}(Y_{10})$ is $F^{\text{UB}}_{Y^N,11}$ and the c.d.f. of $\overline{k}(Y_{10})$ is $F^{\text{LB}}_{Y^N,11}$. The bounds on τ are then

$$\tau_{\text{LB}} = \mathbb{E}[Y_{11}] - \mathbb{E}[\overline{k}(Y_{10})] \quad \text{and} \quad \tau_{\text{UB}} = \mathbb{E}[Y_{11}] - \mathbb{E}[\underline{k}(Y_{10})],$$

with the corresponding estimators

$$\hat{\tau}_{\text{LB}} = \frac{1}{N_{11}} \sum_{i=1}^{N_{11}} Y_{11,i} - \frac{1}{N_{10}} \sum_{i=1}^{N_{10}} \hat{\overline{k}}(Y_{10,i}) \quad \text{and}$$

$$\hat{\tau}_{\text{UB}} = \frac{1}{N_{11}} \sum_{i=1}^{N_{11}} Y_{11,i} - \frac{1}{N_{10}} \sum_{i=1}^{N_{10}} \hat{\underline{k}}(Y_{10,i}).$$

THEOREM 5.4—Asymptotic Distribution for Bounds: *Suppose Assumptions* 5.1(i), (ii), (iv) *and* 5.2 *hold*. *Then*

$$\sqrt{N}(\hat{\tau}_{\text{UB}} - \tau_{\text{UB}}) \xrightarrow{d} \mathcal{N}(0, V^s/\alpha_{11} + \underline{V}^r/\alpha_{10})$$

and

$$\sqrt{N}(\hat{\tau}_{\text{LB}} - \tau_{\text{LB}}) \xrightarrow{d} \mathcal{N}(0, V^s/\alpha_{11} + \overline{V}^r/\alpha_{10}),$$

where $\underline{V}^r = \text{Var}(\underline{k}(Y_{10}))$ *and* $\overline{V}^r = \text{Var}(\overline{k}(Y_{10}))$.

See Appendix A for the proof.

The asymptotic distribution for the bounds can then be used to construct confidence intervals for the parameters of interest, following the work of Imbens and Manski (2004).

Note the difference between the asymptotic variances for the bounds and the variance for the continuous CIC estimator. In the discrete case, the estimation error for the transformations $\underline{k}(\cdot)$ and $\overline{k}(\cdot)$ does not affect the variance of the estimates for the lower and upper bounds. This is because the estimators for $\underline{k}(\cdot)$ and $\overline{k}(\cdot)$ converge to their probability limits faster than $\sqrt{N}$.[34]

5.3. *Inference with Panel Data*

In this section we modify the results to allow for panel data instead of repeated cross sections. Consider first the continuous case. We make the following assumptions regarding the sampling process. Let (Y_{i0}, Y_{i1}) denote the pair of first- and second-period outcomes for unit i.

ASSUMPTION 5.3—Data Generating Process:

(i) *Conditional on $G_i = g$, the pair (Y_{i0}, Y_{i1}) is a random draw from the subpopulation with $G_i = g$.*

(ii) *For $g \in \{0, 1\}$, $\alpha_g \equiv \Pr(G_i = g) > 0$.*

(iii) *The four random variables Y_{gt} are continuous with densities bounded and bounded away from zero with support $\mathbb{Y}_{gt}$ that is a compact subset of $\mathbb{R}$.*

We now have two random samples, one from each group, with sample sizes N_0 and N_1, respectively, and $N = N_0 + N_1$. (In terms of the previous notation, $N_0 = N_{00} = N_{01}$ and $N_1 = N_{10} = N_{11}$.) For each individual we observe Y_{i0} and Y_{i1}. Although we can still linearize the estimator as $\hat{\tau} = \tau + \sum p(Y_{00,i})/N_{00} + \sum q(Y_{01,i})/N_{01} + \sum r(Y_{10,i})/N_{10} + \sum s(Y_{11,i})/N_{11} + o_p(N^{-1/2})$, the four terms in this linearization are no longer independent. The following theorem formalizes the changes in the asymptotic distribution.

THEOREM 5.5—Consistency and Asymptotic Normality: *Suppose Assumption 5.3 holds. Then*:

(i) $\hat{\tau}^{\text{CIC}} \xrightarrow{p} \tau^{\text{CIC}}$;

(ii) $\sqrt{N}(\hat{\tau}^{\text{CIC}} - \tau^{\text{CIC}}) \xrightarrow{d} \mathcal{N}(0, V^p/\alpha_0 + V^q/\alpha_0 + C^{pq}/\alpha_0 + V^r/\alpha_1 + V^s/\alpha_1 + C^{rs}/\alpha_1)$, *where V^p, V^q, V^r, and V^s are as before, and*

$$C^{pq} = \mathbb{E}[p(Y_{00}) \cdot q(Y_{01})] \quad \textit{and}$$

$$C^{rs} = \mathbb{E}[r(Y_{10}) \cdot s(Y_{11})] = \text{Covar}(k(Y_{10}), Y_{11}).$$

See the supplement (Athey and Imbens (2006)) for the proof.

[34] Again a similar situation arises when estimating the median of a discrete distribution. Suppose Z is binary with $\Pr(Z = 1) = p$. The median is $m = \mathbb{1}\{p \geq 1/2\}$ and the estimator is $\hat{m} = \mathbb{1}\{\hat{F}_Z(0) < 1/2\}$. If $p \neq 1/2$, then $\sqrt{N}(\hat{m} - m) \to 0$.

The variances V^p, V^q, V^r, and V^s can be estimated as before. For C^{pq} and C^{rs} we use the estimators

$$\hat{C}^{pq} = \frac{1}{N_0}\sum_{i=1}^{N_0} \hat{p}(Y_{00,i}) \cdot \hat{q}(Y_{01,i}) \quad \text{and} \quad \hat{C}^{rs} = \frac{1}{N_1}\sum_{i=1}^{N_1} \hat{r}(Y_{10,i}) \cdot \hat{s}(Y_{11,i}).$$

THEOREM 5.6—Consistent Estimation of the Variance with Panel Data: *Suppose Assumption* 5.3 *holds and* $\mathbb{Y}_{10} \subseteq \mathbb{Y}_{00}$. *Then* $\hat{V}^p \xrightarrow{p} V^p$, $\hat{V}^q \xrightarrow{p} V^q$, $\hat{V}^r \xrightarrow{p} V^r$, $\hat{V}^s \xrightarrow{p} V^s$, $\hat{C}^{pq} \xrightarrow{p} C^{pq}$, *and* $\hat{C}^{rs} \xrightarrow{p} C^{rs}$.

Now consider the discrete model with panel data.

THEOREM 5.7—Asymptotic Distribution for Bounds: *Suppose Assumptions* 5.2 *and* 5.3(i) *and* (ii) *hold*. *Then*

$$\sqrt{N}(\hat{\tau}_{\text{UB}} - \tau_{\text{UB}}) \xrightarrow{d} \mathcal{N}(0, V^s/\alpha_1 + \underline{V}^r/\alpha_1 + \underline{C}^{rs}/\alpha_1)$$

and

$$\sqrt{N}(\hat{\tau}_{\text{LB}} - \tau_{\text{LB}}) \xrightarrow{d} \mathcal{N}(0, V^s/\alpha_1 + \overline{V}^r/\alpha_{10} + \overline{C}^{rs}/\alpha_1),$$

where $\underline{V}^r = \text{Var}(\underline{k}(Y_{10}))$, $\overline{V}^r = \text{Var}(\overline{k}(Y_{10}))$, $\underline{C}^{rs} = \text{Covar}(\underline{k}(Y_{10}), Y_{11})$, *and* $\overline{C}^{rs} = \text{Covar}(\overline{k}(Y_{10}), Y_{11})$.

See the supplement (Athey and Imbens (2006)) for the proof.

6. MULTIPLE GROUPS AND MULTIPLE TIME PERIODS: IDENTIFICATION, ESTIMATION, AND TESTING

So far we have focused on the simplest setting for DID methods, namely the two-group and two time-period case (from hereon, the 2×2 case). In many applications, however, researchers have data from multiple groups and multiple time periods with different groups receiving the treatment at different times. In this section we discuss the extension of our proposed methods to these cases.[35] We provide large sample results based on a fixed number of groups and time periods. We generalize the assumptions of the CIC model by applying them to all pairs of groups and pairs of time periods. An important feature of the generalized model is that the estimands of interest, e.g., the average effect of the

[35]To avoid repetition, we focus in this section mainly on the average effects of the intervention for the continuous case for the group that received the treatment in the case of repeated cross sections. We can deal with quantile effects, discrete outcomes, effects for the control group, and panel data by generalizing the 2×2 case in an analogous way.

treatment, will differ by group and time period. One reason is that an intrinsic property of our model is that the production function $h(u, t)$ is not restricted as a function of time. Hence even holding the group (the distribution of the unobserved component U) fixed and even if the production function under treatment $h^I(u, t)$ does not vary over time, the average effect of the treatment may vary by time period. Similarly, because the groups differ in their distribution of unobservables, they will differ in the average or quantile effects of the intervention.[36] Initially we therefore focus on estimation of the average treatment effects separately by group and time period.

To estimate the average effect of the intervention for group g in time period t, we require a control group g' and a baseline time period $t' < t$ such that the control group g' is not exposed to the treatment in either of the time periods t and t', and the treatment group g is not exposed to the treatment in the initial time period t'. Under the assumptions of the CIC model, any pair (g', t') that satisfies these conditions will estimate the same average treatment effect. More efficient estimators can be obtained by combining estimators from different control groups and baseline time periods.

The different control groups and different baseline time periods can also be used to test the maintained assumptions of the CIC model. For example, such tests can be used to assess the presence of additive group–period effects. The presence of multiple groups and/or multiple time periods has previously been exploited to construct confidence intervals that are robust to the presence of additive random group–period effects (e.g., Bertrand, Duflo, and Mullainathan (2004), Donald and Lang (2001)). Those results rely critically on the linearity of the estimators to ensure that the presence of such effects does not introduce any bias. As a result, in the current setting the presence of additive group–period effects would in general lead to bias. Moreover, outside of fully parametric models with distributional assumptions, inference in such settings requires large numbers of groups and/or periods even in the linear case.

6.1. *Identification in the Multiple Group and Multiple Time-Period Case*

As before, let $\mathcal{G}$ and $\mathcal{T}$ be the set of group and time indices, where now $\mathcal{G} = \{1, 2, \ldots, N_G\}$ and $\mathcal{T} = \{1, 2, \ldots, N_T\}$. Let $\mathcal{I}$ be the set of pairs (t, g) such that units in period t and group g receive the treatment, with the cardinality of this set equal to $N_{\mathcal{I}}$.[37] For unit i the group indicator is $G_i \in \mathcal{G}$ and the time indicator is $T_i \in \mathcal{T}$. Let I_i be a binary indicator for the treatment received, so that $I_i = 1$ if $(T_i, G_i) \in \mathcal{I}$. We assume that no group receives the treatment in the initial period: $(1, g) \notin \mathcal{I}$. In addition, we assume that after receiving the

[36]This issue of differential effects by group arose already in the discussion of the average effect of the treatment on the treated versus the average effect of the treatment on the control group.

[37]In the 2×2 case, $\mathcal{G} = \{0, 1\}$, $\mathcal{T} = \{0, 1\}$, and $\mathcal{I} = \{(1, 1)\}$ with $N_{\mathcal{I}} = 1$.

treatment, a group continues receiving the treatment in all remaining periods, so that if $t, t+1 \in \mathcal{T}$ and $(t, g) \in \mathcal{I}$, then $(t+1, g) \in \mathcal{I}$. Let $F_{Y,g,t}(y)$ be the distribution function of the outcome in group g and time period t, and let $\alpha_{g,t}$ be the population proportions of each subsample, for $g \in \mathcal{G}$ and $t \in \mathcal{T}$. As before, $Y^N = h(U, t)$ is the production function in the absence of the intervention.

For each "target" pair $(g, t) \in \mathcal{I}$, define the average effect of the intervention:

$$\tau^{\text{CIC}}_{g,t} = \mathbb{E}[Y^I_{g,t} - Y^N_{g,t}] = \mathbb{E}[Y^I_{g,t}] - \mathbb{E}[h(U, t)|G = g].$$

This average treatment effect potentially differs by target group–period (g, t) because we restrict neither the distribution of Y^I by group and time nor the production function $h(u, t)$ beyond monotonicity in the unobserved component.

In the 2×2 case there was a single control group and a single baseline time period. Here $\tau^{\text{CIC}}_{g,t}$ can be estimated in a number of different ways, using a range of control groups and baseline time periods. Formally, we can use any control group $g_0 \neq g$ in time period $t_0 < t$ as long as $(g_0, t_0), (g_0, t), (g, t_0) \notin \mathcal{I}$. It is useful to introduce a separate notation for these objects. For each (g, t), which defines the target group g and time period t, and for each control group and baseline time period (g_0, t_0), define

$$\kappa_{g_0,g,t_0,t} = \mathbb{E}[Y_{g,t}] - \mathbb{E}\big[F^{-1}_{Y,g_0,t}(F_{Y,g_0,t_0}(Y_{g,t_0}))\big].$$

As before, the identification question concerns conditions under which $\mathbb{E}[F^{-1}_{Y,g_0,t}(F_{Y,g_0,t_0}(Y_{g,t_0}))] = \mathbb{E}[Y^N_{g,t}]$, implying $\kappa_{g_0,g,t_0,t} = \tau^{\text{CIC}}_{g,t}$. Here we present a generalization of Theorem 3.1. For ease of exposition, we strengthen the support assumption, although this can be relaxed as in the 2×2 case.

ASSUMPTION 6.1—Support in the Multiple Group and Multiple Time-Period Case: *The support of $U|G = g$, denoted by $\mathbb{U}_g$, is the same for all $g \in \mathcal{G}$.*

THEOREM 6.1—Identification in the Multiple Group and Multiple Time-Period Case: *Suppose Assumptions* 3.1–3.3 *and* 6.1 *hold. Then for any (g_1, t_1) with $(g_1, t_1) \in \mathcal{I}$ such that there is a pair (g_0, t_0) that satisfies $(g_0, t_0), (g_0, t_1), (g_1, t_0) \notin \mathcal{I}$, the distribution of $Y^N_{g_1,t_1}$ is identified and, for any such (g_0, t_0),*

$$F_{Y^N,g_1,t_1}(y) = F_{Y,g_1,t_0}\big(F^{-1}_{Y,g_0,t_0}(F_{Y,g_0,t_1}(y))\big). \tag{45}$$

The proof of Theorem 6.1 is similar to that of Theorem 3.1 and is omitted.

The implication of this theorem is that for all control groups and baseline time periods (g_0, t_0) that satisfy the conditions in Theorem 6.1, we have $\tau^{\text{CIC}}_{g_1,t_1} = \kappa_{g_0,g_1,t_0,t_1}$.

6.2. *Inference in the Multiple Group and Multiple Time-Period Case*

The focus of this section is estimation of and inference for $\tau^{\text{CIC}}_{g,t}$. As a first step, we consider inference for κ_{g_0,g_1,t_0,t_1}. For each quadruple (g_0, g_1, t_0, t_1), we can estimate the corresponding κ_{g_0,g_1,t_0,t_1} as

$$\hat{\kappa}_{g_0,g_1,t_0,t_1} = \frac{1}{N_{g_1,t_1}} \sum_{i=1}^{N_{g_1,t_1}} Y_{g_1,t_1,i} - \frac{1}{N_{g_1,t_0}} \sum_{i=1}^{N_{g_1,t_0}} \hat{F}^{-1}_{Y,g_0,t_1}(\hat{F}_{Y,g_0,t_0}(Y_{g_1,t_0,i})). \tag{46}$$

By Theorem 6.1, if $t_0 < t_1$, $(g_1, t_1) \in \mathcal{I}$, and $(g_0, t_0), (g_0, t_1), (g_1, t_0) \notin \mathcal{I}$, it follows that $\kappa_{g_0,g_1,t_0,t_1} = \tau^{\text{CIC}}_{g_1,t_1}$. Hence we have potentially many consistent estimators for each $\tau^{\text{CIC}}_{g,t}$. Here we first analyze the properties of each $\hat{\kappa}_{g_0,g_1,t_0,t_1}$ as an estimator for κ_{g_0,g_1,t_0,t_1}, and then consider combining the different estimators into a single estimator $\hat{\tau}_{g,t}$ for $\tau_{g,t}$.

For inference concerning κ_{g_0,g_1,t_0,t_1}, we exploit the asymptotic linearity of the estimators $\hat{\kappa}_{g_0,g_1,t_0,t_1}$. To do so it is useful to index the previously defined functions $p(\cdot)$, $q(\cdot)$, $r(\cdot)$, and $s(\cdot)$ by groups and time periods. First, define[38]

$$P_{g_0,g_1,t_0,t_1}(y,z) = \frac{1}{f_{Y,g_0,t_1}(F^{-1}_{Y,g_0,t_1}(F_{Y,g_0,t_0}(z)))} \cdot (\mathbb{1}\{y \le z\} - F_{Y,g_0,t_0}(z)),$$

$$Q_{g_0,g_1,t_0,t_1}(y,z) = -\frac{1}{f_{Y,g_0,t_1}(F^{-1}_{Y,g_0,t_1}(F_{Y,g_0,t_0}(z)))} \times \big(\mathbb{1}\{F_{Y,g_0,t_1}(y) \le F_{Y,g_0,t_0}(z)\} - F_{Y,g_0,t_0}(z)\big),$$

$$p_{g_0,g_1,t_0,t_1}(y) = \mathbb{E}[P_{g_0,g_1,t_0,t_1}(y, Y_{g_1,t_0})],$$

$$q_{g_0,g_1,t_0,t_1}(y) = \mathbb{E}[Q_{g_0,g_1,t_0,t_1}(y, Y_{g_1,t_0})],$$

$$r_{g_0,g_1,t_0,t_1}(y) = F^{-1}_{Y,g_0,t_1}(F_{Y,g_0,t_0}(y)) - \mathbb{E}\big[F^{-1}_{Y,g_0,t_1}(F_{Y,g_0,t_0}(Y_{g_1,t_0}))\big],$$

and

$$s_{g_0,g_1,t_0,t_1}(y) = y - \mathbb{E}[Y_{g_1,t_1}].$$

Also define the four averages

$$\hat{\mu}^p_{g_0,g_1,t_0,g_1} = \frac{1}{N_{g_0,t_0}} \sum_{i=1}^{N_{g_0,t_0}} p_{g_0,g_1,t_0,t_1}(Y_{g_0,t_0,i}),$$

$$\hat{\mu}^q_{g_0,g_1,t_0,g_1} = \frac{1}{N_{g_0,t_1}} \sum_{i=1}^{N_{g_0,t_1}} q_{g_0,g_1,t_0,t_1}(Y_{g_0,t_1,i}),$$

[38]Although we index the function $s_{g_0,g_1,t_0,t_1}(y)$ by g_0, g_1, t_0, and t_1 only to make it comparable to the others, it does not actually depend on group or time.

$$\hat{\mu}^r_{g_0,g_1,t_0,g_1} = \frac{1}{N_{g_1,t_0}} \sum_{i=1}^{N_{g_1,t_0}} r_{g_0,g_1,t_0,t_1}(Y_{g_1,t_0,i}),$$

$$\hat{\mu}^s_{g_0,g_1,t_0,g_1} = \frac{1}{N_{g_1,t_1}} \sum_{i=1}^{N_{g_1,t_1}} s_{g_0,g_1,t_0,t_1}(Y_{g_1,t_1,i}).$$

Define the normalized variances of the $\hat{\mu}$'s:

$$V^p_{g_0,g_1,t_0,t_1} = N_{g_0,t_0} \cdot \mathrm{Var}(\hat{\mu}^p_{g_0,g_1,t_0,g_1}),$$
$$V^q_{g_0,g_1,t_0,t_1} = N_{g_0,t_1} \cdot \mathrm{Var}(\hat{\mu}^q_{g_0,g_1,t_0,g_1}),$$
$$V^r_{g_0,g_1,t_0,t_1} = N_{g_1,t_0} \cdot \mathrm{Var}(\hat{\mu}^r_{g_0,g_1,t_0,g_1}),$$
$$V^s_{g_0,g_1,t_0,t_1} = N_{g_1,t_1} \cdot \mathrm{Var}(\hat{\mu}^s_{g_0,g_1,t_0,g_1}).$$

Finally, define

$$\tilde{\kappa}_{g_0,g_1,t_0,t_1} = \kappa_{g_0,g_1,t_0,t_1} + \hat{\mu}^p_{g_0,g_1,t_0,g_1} + \hat{\mu}^q_{g_0,g_1,t_0,g_1} + \hat{\mu}^r_{g_0,g_1,t_0,g_1} + \hat{\mu}^s_{g_0,g_1,t_0,g_1}.$$

LEMMA 6.1—Asymptotic Linearity: *Suppose Assumptions* 5.1 *and* 6.1 *hold. Then* $\hat{\kappa}_{g_0,g_1,t_0,t_1}$ *is asymptotically linear*: $\hat{\kappa}_{g_0,g_1,t_0,t_1} = \tilde{\kappa}_{g_0,g_1,t_0,t_1} + o_p(N^{-1/2})$.

The proof of Lemma 6.1 follows directly from that of Theorem 5.1.

The implication of this lemma is that the normalized asymptotic variance of $\hat{\kappa}^{\mathrm{CIC}}_{g_0,g_1,t_0,t_1}$ is equal to the normalized variance of $\tilde{\kappa}^{\mathrm{CIC}}_{g_0,g_1,t_0,t_1}$, which is equal to

$$N \cdot \mathrm{Var}(\tilde{\kappa}_{g_0,g_1,t_0,t_1}) = \frac{V^p_{g_0,g_1,t_0,t_1}}{\alpha_{g_0,t_0}} + \frac{V^q_{g_0,g_1,t_0,t_1}}{\alpha_{g_0,t_1}} + \frac{V^r_{g_0,g_1,t_0,t_1}}{\alpha_{g_1,t_0}} + \frac{V^s_{g_0,g_1,t_0,t_1}}{\alpha_{g_1,t_1}}.$$

In addition to the variance, we also need the normalized large sample covariance between $\hat{\kappa}_{g_0,g_1,t_0,t_1}$ and $\hat{\kappa}_{g_0',g_1',t_0',t_1'}$. There are 25 cases (including the case with $g_0 = g_0'$, $g_1 = g_1'$, $t_0 = t_0'$, and $t_1 = t_1'$, where the covariance is equal to the variance). For example, if $g_0 = g_0'$, $g_1 = g_1'$, $t_0 = t_0'$, and $t_1 \neq t_1'$, then the normalized covariance is

$$\begin{aligned} &N \cdot \mathrm{Cov}(\tilde{\kappa}^{\mathrm{CIC}}_{g_0,g_1,t_0,t_1}, \tilde{\kappa}^{\mathrm{CIC}}_{g_0',g_1',t_0',t_1'}) \\ &\quad = N \cdot \mathrm{Cov}(\tilde{\kappa}^{\mathrm{CIC}}_{g_0,g_1,t_0,t_1}, \tilde{\kappa}^{\mathrm{CIC}}_{g_0,g_1,t_0,t_1'}) \\ &\quad = N \cdot \mathbb{E}[\hat{\mu}^p_{g_0,g_1,t_0,t_1} \cdot \hat{\mu}^p_{g_0,g_1,t_0,t_1'}] + N \cdot \mathbb{E}[\hat{\mu}^r_{g_0,g_1,t_0,t_1} \cdot \hat{\mu}^r_{g_0,g_1,t_0,t_1'}]. \end{aligned}$$

The details of the full set of 25 cases are given in Appendix B.

Let $\mathcal{J}$ be the set of quadruples (g_0, g_1, t_0, t_1) such that $(g_0, t_0), (g_0, t_1), (g_1, t_0) \notin \mathcal{I}$ and $(g_1, t_1) \in \mathcal{I}$, and let $N_{\mathcal{J}}$ be the cardinality of this set. Stack

all $\hat{\kappa}_{g_0,g_1,t_0,t_1}$ such that $(g_0, g_1, t_0, t_1) \in \mathcal{J}$ into the $N_{\mathcal{J}}$-dimensional vector $\hat{\kappa}_{\mathcal{J}}$; similarly stack the κ_{g_0,g_1,t_0,t_1} into the $N_{\mathcal{J}}$-dimensional vector $\kappa_{\mathcal{J}}$. Let $V_{\mathcal{J}}$ be the asymptotic covariance matrix of $\sqrt{N} \cdot \hat{\kappa}_{\mathcal{J}}$.

THEOREM 6.2: *Suppose Assumptions* 5.1 *and* 6.1 *hold. Then*

$$\sqrt{N}(\hat{\kappa}_{\mathcal{J}} - \kappa_{\mathcal{J}}) \xrightarrow{d} \mathcal{N}(0, V_{\mathcal{J}}).$$

For the proof, see Appendix A.

Next, we wish to combine the different estimates of $\tau^{\text{CIC}}_{g,t}$. To do so efficiently, we need to estimate the covariance matrix of the estimators $\hat{\kappa}_{g_0,g_1,t_0,t_1}$, $V_{\mathcal{J}}$. As shown in Appendix A, all the covariance terms involve expectations of products of the functions $p_{g_0,g_1,t_0,t_1}(y)$, $q_{g_0,g_1,t_0,t_1}(y)$, $r_{g_0,g_1,t_0,t_1}(y)$, and $s_{g_0,g_1,t_0,t_1}(y)$, evaluated over the distribution of $Y_{g,t}$. These expectations can be estimated by averaging over the sample. Let the resulting estimator for $V_{\mathcal{J}}$ be denoted by $\hat{V}_{\mathcal{J}}$. The following lemma, implied by Theorem 5.2, states its consistency.

LEMMA 6.2: *Suppose Assumption* 5.1 *holds. Then* $\hat{V}_{\mathcal{J}} \xrightarrow{p} V_{\mathcal{J}}$.

It is important to note that the covariance matrix $V_{\mathcal{J}}$ is not necessarily of full rank.[39] In that case we denote the (Moore–Penrose) generalized inverse of the matrix $V_{\mathcal{J}}$ by $V_{\mathcal{J}}^{(-)}$.

We wish to combine the estimators for κ_{g_0,g_1,t_0,t_1} into estimators for $\tau^{\text{CIC}}_{g,t}$. Let $\tau^{\text{CIC}}_{\mathcal{I}}$ denote the vector of length $N_{\mathcal{I}}$ that consists of all $\tau^{\text{CIC}}_{g,t}$ stacked. In addition, let A denote the $N_{\mathcal{J}} \times N_{\mathcal{I}}$ matrix of 0–1 indicators such that $\kappa_{\mathcal{J}} = A \cdot \tau^{\text{CIC}}_{\mathcal{I}}$ under the assumptions of Theorem 6.1. Specifically, under the assumptions of Theorem 6.1, if the jth element of $\kappa_{\mathcal{J}}$ is equal to the ith element of $\tau^{\text{CIC}}_{\mathcal{I}}$, then (i, j)th element of A is equal to 1. Then we estimate $\tau^{\text{CIC}}_{\mathcal{I}}$ as

$$\hat{\tau}^{\text{CIC}}_{\mathcal{I}} = (A'\hat{V}_{\mathcal{J}}^{(-)}A)^{-1}(A'\hat{V}_{\mathcal{J}}^{(-)}\hat{\kappa}^{\text{CIC}}_{\mathcal{J}}).$$

[39]To see how this may arise, consider a simple example with four groups ($\mathcal{G} = \{1, 2, 3, 4\}$) and two time periods ($\mathcal{T} = \{1, 2\}$). Suppose only the last two groups (groups 3 and 4) receive the treatment in the second period, so that $(3, 2), (4, 2) \in \mathcal{I}$ and all other combinations of $(g, t) \notin \mathcal{I}$. There are two treatment effects—$\tau^{\text{CIC}}_{3,2}$ and $\tau^{\text{CIC}}_{4,2}$—and four comparisons that estimate these two treatment effects—$\kappa_{1,3,1,2}$ and $\kappa_{2,3,1,2}$, which are both equal to $\tau^{\text{CIC}}_{3,2}$, and $\kappa_{1,4,1,2}$ and $\kappa_{2,4,1,2}$, which are both equal to $\tau^{\text{CIC}}_{4,2}$. Suppose also that $F_{Y,g,t}(y) = y$ for all g, t. In that case, simple calculations show $\mathbb{E}[p_{g_0,g_1,t_0,t_1}(y)] = \mathbb{E}[q_{g_0,g_1,t_0,t_1}(y)] = r_{g_0,g_1,t_0,t_1}(y) = s_{g_0,g_1,t_0,t_1}(y) = y - 1/2$, so that $\tilde{\kappa}_{1,3,1,2} = \bar{Y}_{3,2} - \bar{Y}_{3,1} - \bar{Y}_{1,2} - \bar{Y}_{1,1}$, $\tilde{\kappa}_{1,4,1,2} = \bar{Y}_{4,2} - \bar{Y}_{4,1} - \bar{Y}_{1,2} - \bar{Y}_{1,1}$, $\tilde{\kappa}_{2,3,1,2} = \bar{Y}_{3,2} - \bar{Y}_{3,1} - \bar{Y}_{2,2} - \bar{Y}_{2,1}$, and $\tilde{\kappa}_{2,4,1,2} = \bar{Y}_{4,2} - \bar{Y}_{4,1} - \bar{Y}_{2,2} - \bar{Y}_{2,1}$. Then $\tilde{\kappa}_{2,4,1,2} - \tilde{\kappa}_{2,3,1,2} - \tilde{\kappa}_{1,4,1,2} + \tilde{\kappa}_{1,3,1,2} = 0$, which shows that the covariance matrix of the four estimators is asymptotically singular. In general, the covariance matrix will have full rank, but we need to allow for special cases such as these.

THEOREM 6.3: *Suppose Assumptions* 3.1–3.3, 5.1, *and* 6.1 *hold*. *Then*

$$\sqrt{N} \cdot (\hat{\tau}^{\text{CIC}}_{\mathcal{I}} - \tau^{\text{CIC}}_{\mathcal{I}}) \xrightarrow{d} \mathcal{N}\big(0, (A' V^{(-)}_{\mathcal{J}} A)^{-1}\big).$$

PROOF: A linear combination of a jointly normal random vector is normally distributed. The mean and variance then follow directly from those for $\hat{\kappa}_{\mathcal{J}}$. *Q.E.D.*

In some cases we may wish to combine these estimates further. For example, suppose we may wish to estimate a single effect for a particular group, combining estimates for all periods in which this group was exposed to the intervention. Alternatively, we may be interested in estimating a single effect for each time period, combining all estimates from groups exposed to the intervention during that period. We may even wish to combine estimates for different groups and periods into a single average estimate of the effect of the intervention. In general, we can consider estimands of the form $\tau^{\text{CIC}}_{\Lambda} = \Lambda' \tau^{\text{CIC}}_{\mathcal{I}}$, where Λ is an $N_{\mathcal{I}} \times L$ matrix of weights with each column adding up to 1. If we are interested in a single average, $L = 1$; more generally, we may be interested in a vector of effects, e.g., one for each group or each time period. The weights may be choosen to reflect relative sample sizes or to depend on the variances of the $\hat{\tau}^{\text{CIC}}_{\mathcal{I}}$. The natural estimator for $\tau^{\text{CIC}}_{\Lambda}$ is $\hat{\tau}^{\text{CIC}}_{\Lambda} = \Lambda' \hat{\tau}^{\text{CIC}}_{\mathcal{I}}$. For fixed Λ it satisfies

$$\sqrt{N} \cdot (\hat{\tau}^{\text{CIC}}_{\Lambda} - \tau^{\text{CIC}}_{\Lambda}) \xrightarrow{d} \mathcal{N}\big(0, \Lambda' (A' V^{(-)}_{\mathcal{J}} A)^{-1} \Lambda\big).$$

As an example, suppose one wishes to estimate a single average effect, so Λ is an $N_{\mathcal{I}}$ vector and (with some abuse of notation) $\tau^{\text{CIC}}_{\Lambda} = \sum_{(g,t)\in\mathcal{I}} \Lambda_{g,t} \cdot \tau^{\text{CIC}}_{g,t}$. One natural choice is to weight by the sample sizes of the group–time periods, so $\Lambda_{g,t} = N_{g,t} / \sum_{(g,t)\in\mathcal{I}} N_{g,t}$. Alternatively, one can weight using the variances, leading to $\Lambda = (\iota' A'_{\mathcal{I}} V^{(-)}_{\mathcal{J}} A \iota)^{-1} \iota' A' V^{(-)}_{\mathcal{J}} A$. This latter choice is particularly appropriate under the (strong) assumption that the treatment effect does not vary by group or time period, although the above large sample results do not require this assumption.

6.3. *Testing*

In addition to combining the vector of estimators to obtain a more efficient estimator for τ^{CIC}, we can also use it to test the assumptions of the CIC model. Under the maintained assumptions, all estimates of the form $\hat{\kappa}_{g_0,g_1,t_0,t_1}$ will estimate $\tau^{\text{CIC}}_{g_1,t_1}$. If the model is misspecified, the separate estimators may converge to different limiting values. We can implement this test as follows.

THEOREM 6.4: *Suppose that Assumptions* 3.1–3.3, 5.1, *and* 6.1 *hold*. *Then*

$$N \cdot (\hat{\kappa}_{\mathcal{J}} - A \cdot \hat{\tau}^{\text{CIC}}_{\mathcal{I}})' \hat{V}^{(-)}_{\mathcal{J}} (\hat{\kappa}_{\mathcal{J}} - A \cdot \hat{\tau}^{\text{CIC}}_{\mathcal{I}}) \xrightarrow{d} \mathcal{X}^2(\text{rank}(V_{\mathcal{J}}) - N_{\mathcal{I}}).$$

PROOF: By joint normality of $\hat{\kappa}_{\mathcal{J}}$ and the definition of $\hat{\tau}^{\text{CIC}}_{\mathcal{I}}$, it follows that $\hat{\kappa}_{\mathcal{J}} - A \cdot \hat{\tau}^{\text{CIC}}_{\mathcal{I}}$ is jointly normal with mean zero and covariance matrix with $\text{rank}(V_{\mathcal{J}}) - N_{\mathcal{I}}$. *Q.E.D.*

This test will have power against a number of violations of the assumptions. In particular, it will have power against violations of the assumption that the unobserved component is independent of the time period conditional on the group or $U \perp T|G$. One form such violations could take is through additive random group–time effects. In additive linear DID models such random group–time effects do not introduce bias, although, for inference, the researcher relies either on distributional assumptions or on asymptotics based on large numbers of groups or time periods (e.g., Bertrand, Duflo, and Mullainathan (2004), Donald and Lang (2001)). In the current setting, the presence of such effects can introduce bias because of the nonadditivity and nonlinearity of $h(u, t)$. There appears to be no simple adjustment to remove this bias. Fortunately, the presence of such effects is testable using Theorem 6.4.

We may wish to further test equality of $\tau^{\text{CIC}}_{g,t}$ for different g and t. Such tests can be based on the same approach as used in Theorem 6.4. As an example, consider testing the null hypothesis that $\tau^{\text{CIC}}_{g,t} = \tau^{\text{CIC}}$ for all $(g, t) \in \mathcal{I}$. In that case, we first estimate τ^{CIC} as $\hat{\tau}^{\text{CIC}} = \Lambda \hat{\tau}^{\text{CIC}}_{\mathcal{I}}$ with $\Lambda = (\iota' A'_{\mathcal{I}} V^{(-)}_{\mathcal{J}} A \iota)^{-1} \iota' A' V^{(-)}_{\mathcal{J}} A$. Then the test statistic is $N \cdot (\hat{\tau}^{\text{CIC}}_{\mathcal{I}} - \hat{\tau}^{\text{CIC}} \cdot \iota)' A'_{\mathcal{I}} V^{(-)}_{\mathcal{J}} A (\hat{\tau}^{\text{CIC}}_{\mathcal{I}} - \hat{\tau}^{\text{CIC}} \cdot \iota)$. In large samples, $N \cdot (\hat{\tau}^{\text{CIC}}_{\mathcal{I}} - \hat{\tau}^{\text{CIC}} \cdot \iota)' A'_{\mathcal{I}} V^{(-)}_{\mathcal{J}} A (\hat{\tau}^{\text{CIC}}_{\mathcal{I}} - \hat{\tau}^{\text{CIC}} \cdot \iota) \xrightarrow{d} \mathcal{X}^2(N_{\mathcal{I}} - 1)$ under the null hypothesis of $\tau^{\text{CIC}}_{g,t} = \tau^{\text{CIC}}$ for all groups and time periods.

7. CONCLUSION

In this paper, we develop a new approach to difference-in-differences models that highlights the role of changes in entire distribution functions over time. Using our methods, it is possible to evaluate a range of economic questions suggested by policy analysis, such as questions about mean–variance trade-offs or which parts of the distribution benefit most from a policy, while maintaining a single, internally consistent economic model of outcomes.

The model we focus on, the changes-in-changes model, has several advantages. It is considerably more general than the standard DID model. Its assumptions are invariant to monotone transformations of the outcome. It allows the distribution of unobservables to vary across groups in arbitrary ways. For example, it allows for the possibility that the distribution of outcomes in the absence of the policy intervention would change over time in both mean and variance. Our method could evaluate the effects of a policy on the mean and variance of the treatment group's distribution relative to the underlying time trend in these moments.

A number of issues concerning DID methods have been debated in the literature. One common concern (e.g., Besley and Case (2000)) is that the effects

identified by DID may not have a causal interpretation if the policy change occurred in a jurisdiction that derives unusual benefits from the policy change. That is, the treatment group may differ from the control group in the effects of the treatment, not just in terms of the distribution of outcomes in the absence of the treatment. Our approach allows for both of these types of differences across groups because we allow the effect of the treatment to vary by unobservable characteristics whose distribution may vary across groups. As long as there are no differences across groups in the underlying treatment and nontreatment "production functions" that map unobservables to outcomes at a point in time, our approach can provide consistent estimates of the effect of the policy on both the treatment and the control group.

In the supplement for this paper (Athey and Imbens (2006)), we present an application to the problem of disability insurance (Meyer, Viscusi, and Dubin (1995)) that illustrates that our approach to estimate the effects of a policy change can lead to results that differ from those obtained through the standard DID approach in magnitude and significance. Thus, the restrictive assumptions required for standard DID methods can have significant policy implications. Even when one applies the more general classes of models proposed in this paper, however, it will be important to justify such assumptions carefully.

Dept. of Economics, Stanford University, Stanford, CA 94305-6072, U.S.A., and National Bureau of Economic Research; athey@stanford.edu; http://www.stanford.edu/~athey/

and

Dept. of Economics and Dept. of Agricultural and Resource Economics, University of California Berkeley, Berkeley, CA 94720-3880, U.S.A., and National Bureau of Economic Research; imbens@econ.berkeley.edu; http://elsa.berkeley.edu/users/imbens/.

Manuscript received May, 2002; final revision received April, 2005.

APPENDIX A: PROOFS

Before presenting a proof of Theorem 5.1, we give a couple of preliminary results. These results will be used in the construction of an asymptotically linear representation of $\hat{\tau}^{\text{CIC}}$, following the general structure of such proofs for asymptotic normality of semiparametric estimators in Newey (1994). The technical issues involve checking that the asymptotic linearization of $\hat{F}^{-1}_{Y,01}(\hat{F}_{Y,00}(z))$ is uniform in z at the appropriate rate, because $\hat{\tau}^{\text{CIC}}$ involves the average $(1/N_{10})\sum_i \hat{F}^{-1}_{Y,01}(\hat{F}_{Y,00}(Y_{10,i}))$. This in turn will hinge on an asymptotically linear representation of $F^{-1}_{Y,gt}(q)$ that is uniform in $q \in [0,1]$ at the appropriate rate (Lemma A.6). The key result uses a result by Stute (1982), restated here as Lemma A.4, that bounds the supremum of the difference in empirical distribution functions evaluated at points close together. In the Appendix, the ab-

breviations TI and MVT will be used as shorthand for the triangle inequality and the mean value theorem, respectively.

Because $N_{gt}/N \to \alpha_{gt}$, with α_{gt} positive, any term that is $O_p(N_{gt}^{-\delta})$ is also $O_p(N^{-\delta})$; similarly, terms that are $o_p(N_{gt}^{-\delta})$ are $o_p(N^{-\delta})$. In the following discussion for notational convenience we drop the subscript gt when the results are valid for Y_{gt} for all $(g,t) \in \{(0,0),(0,1),(1,0)\}$.

Recall that as an estimator for the distribution function, we use the empirical distribution function

$$\hat{F}_Y(y) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\{Y_j \le y\}$$

$$= F_Y(y) + \frac{1}{N}\sum_{i=1}^{N}\big(\mathbb{1}\{Y_i \le y\} - F_Y(y)\big)$$

and as an estimator of its inverse, we use

$$\text{(A.1)} \qquad \hat{F}_Y^{-1}(q) = Y_{([N\cdot q])} = \inf\{y \in \mathbb{Y} : \hat{F}_Y(y) \ge q\}$$

for $q \in [0,1]$, where $Y_{(k)}$ is the kth order statistic of $Y_1,\ldots,Y_N$ and $[a]$ is the smallest integer greater than or equal to a, so that $F_Y^{-1}(0) = \underline{y}$. Note that

$$\text{(A.2)} \qquad q \le \hat{F}_Y(\hat{F}_Y^{-1}(q)) < q + 1/N,$$

with $\hat{F}_Y(\hat{F}_Y^{-1}(q)) = q$ if $q = j/N$ for some integer $j \in \{0,1,\ldots,N\}$. Also

$$y - \max_{i=1,\ldots,N}(Y_{(i)} - Y_{(i-1)}) < \hat{F}_Y^{-1}(\hat{F}_Y(y)) \le y,$$

where $Y_{(0)} = \underline{y}$, with $\hat{F}_Y^{-1}(\hat{F}_Y(y)) = y$ at all sample values $Y_1,\ldots,Y_N$.

LEMMA A.1: *Let* $\mathbb{U} = [\underline{u}, \overline{u}]$, *let* $\mathbb{Y} = [\underline{y}, \overline{y}]$ *with* $-\infty < \underline{u}, \overline{u}, \underline{y}, \overline{y} < \infty$, *and let* $g(\cdot) : \mathbb{Y} \to \mathbb{U}$ *be a nondecreasing, right continuous function with its inverse defined as*

$$g^{-1}(u) = \inf\{y \in \mathbb{Y} : g(y) \ge u\}.$$

Then:

(i) *For all* $u \in \mathbb{U}$, $g(g^{-1}(u)) \ge u$.
(ii) *For all* $y \in \mathbb{Y}$, $g^{-1}(g(y)) \le y$.
(iii) *For all* $y \in \mathbb{Y}$, $g(g^{-1}(g(y))) = g(y)$.
(iv) *For all* $u \in \mathbb{U}$, $g^{-1}(g(g^{-1}(u))) = g^{-1}(u)$.
(v) *We have* $\{(u,y) | u \in \mathbb{U}, y \in \mathbb{Y}, u \le g(y)\} = \{(u,y) | u \in \mathbb{U}, y \in \mathbb{Y}, g^{-1}(u) \le y\}$.

See the supplement (Athey and Imbens (2006)) for the proof. Note that this lemma applies to the case where $g(y)$ is an (estimated) cumulative distribution function and $g^{-1}(u)$ is the inverse distribution function defined in (A.1).

Next we state a general result regarding the uniform convergence of the empirical distribution function.

LEMMA A.2: *For any* $\delta < 1/2$,

$$\sup_{y\in\mathbb{Y}} N^{\delta}\cdot|\hat{F}_Y(y)-F_Y(y)|\xrightarrow{p}0.$$

PROOF: Billingsley (1968) and Shorack and Wellner (1986) show that with $X_1, X_2, \ldots$ independent and identically distributed, and uniform on $[0,1]$, $\sup_{0\le x\le 1} N^{1/2}\cdot|\hat{F}_X(x)-x| = O_p(1)$. Hence for all $\delta < 1/2$, we have $\sup_{0\le x\le 1} N^{\delta}\cdot|\hat{F}_X(x)-x|\xrightarrow{p}0$. Consider the one-to-one transformation from $\mathbb{X}$ to $\mathbb{Y}$, $Y = F_Y^{-1}(X)$, so that the distribution function for Y is $F_Y(y)$. Then

$$\begin{aligned}\sup_{y\in\mathbb{Y}} N^{\delta}\cdot|\hat{F}_Y(y)-F_Y(y)| &= \sup_{0\le x\le 1} N^{\delta}\cdot\left|\hat{F}_Y(F_Y^{-1}(x))-F_Y(F_Y^{-1}(x))\right| \\ &= \sup_{0\le x\le 1} N^{\delta}\cdot|\hat{F}_X(x)-x|\xrightarrow{p}0,\end{aligned}$$

because

$$\begin{aligned}\hat{F}_X(x) &= (1/N)\sum \mathbb{1}\{F_Y(Y_i)\le x\} \\ &= (1/N)\sum \mathbb{1}\{Y_i\le F_Y^{-1}(x)\} = \hat{F}_Y(F_Y^{-1}(x)).\end{aligned}$$

Q.E.D.

Next, we show that the inverse of the empirical distribution converges at the same rate:

LEMMA A.3: *For any* $\delta < 1/2$,

$$\sup_{q\in[0,1]} N^{\delta}\cdot|\hat{F}_Y^{-1}(q)-F_Y^{-1}(q)|\xrightarrow{p}0.$$

Before proving Lemma A.3 we prove some other results.

Next we state a result concerning uniform convergence of the difference between the difference of the empirical distribution function and its population counterpart and the same difference at a nearby point. The following lemma is for uniform distributions on [0, 1].

LEMMA A.4—Stute (1982): *Let*

$$\begin{aligned}\omega(a) = \sup_{0\le y\le 1, 0\le x\le a, 0\le x+y\le 1} N^{1/2}\cdot\big|&\hat{F}_Y(y+x)-\hat{F}_Y(x) \\ &-(F_Y(y+x)-F_Y(y))\big|.\end{aligned}$$

Suppose that (i) $a_N \to 0$, (ii) $N \cdot a_N \to \infty$, (iii) $\log(1/a_N)/\log\log N \to \infty$, *and* (iv) $\log(1/a_N)/(N \cdot a_N) \to 0$. *Then*

$$\lim_{N\to\infty} \frac{\omega(a_N)}{\sqrt{2a_N \log(1/a_N)}} = 1 \quad w.p.1.$$

For the proof, see Stute (1982, Theorem 0.2) or Shorack and Wellner (1986, Chapter 14.2, Theorem 1).

Using the same argument as in Lemma A.2, one can show that the rate at which $\omega(a)$ converges to zero as a function of a does not change if one relaxes the uniform distribution assumption to allow for a distribution with compact support and continuous density bounded and bounded away from zero. Here we state this in a slightly different way.

LEMMA A.5—Uniform Convergence: *Suppose Assumption* 5.1 *holds*. *Then*, *for* $0 < \eta < 3/4$ *and* $\delta > \max(2\eta - 1, \eta/2)$,

$$\sup_{y\in\mathbb{Y}, x\le N^{-\delta}, x+y\in\mathbb{Y}} N^{\eta} \cdot \left|\hat{F}_Y(y+x) - \hat{F}_Y(y) - (F_Y(y+x) - F_Y(y))\right| \xrightarrow{p} 0.$$

The proof is given in the supplement.

Next we state a result regarding asymptotic linearity of quantile estimators and we provide a rate on the error of this approximation.

LEMMA A.6: *For all* $0 < \eta < 5/7$,

$$\sup_{q\in[0,1]} N^{\eta} \cdot \left|\hat{F}_Y^{-1}(q) - F_Y^{-1}(q) + \frac{1}{f_Y(F_Y^{-1}(q))}\big(\hat{F}_Y(F_Y^{-1}(q)) - q\big)\right| \xrightarrow{p} 0.$$

The proof is given in the supplement.

PROOF OF LEMMA A.3: By the TI,

$$\sup_{q\in[0,1]} N^{\delta} \cdot |\hat{F}_Y^{-1}(q) - F_Y^{-1}(q)|$$

$$\le \sup_{q\in[0,1]} N^{\delta} \cdot \left|\hat{F}_Y^{-1}(q) - F_Y^{-1}(q) + \frac{1}{f_Y(F_Y^{-1}(q))}\big(\hat{F}_Y(F_Y^{-1}(q)) - q\big)\right| \tag{A.3}$$

$$+ \sup_{q\in[0,1]} N^{\delta} \cdot \left|\frac{1}{f_Y(F_Y^{-1}(q))}\big(\hat{F}_Y(F_Y^{-1}(q)) - q\big)\right|. \tag{A.4}$$

By Lemma A.6, (A.3) converges to zero. Next, consider (A.4):

$$\begin{aligned}
&\sup_{q\in[0,1]} N^{\delta}\cdot\left|\frac{1}{f_Y(F_Y^{-1}(q))}\big(\hat{F}_Y(F_Y^{-1}(q))-q\big)\right| \\
&\quad\le \frac{1}{\underline{f}}\sup_{q\in[0,1]} N^{\delta}\cdot\big|\hat{F}_Y(F_Y^{-1}(q))-F_Y(F_Y^{-1}(q))\big| \\
&\quad\le \frac{1}{\underline{f}}\sup_{y\in\mathbb{Y}} N^{\delta}\cdot|\hat{F}_Y(y)-F_Y(y)|,
\end{aligned}$$

which converges to zero by Lemma A.2. *Q.E.D.*

Using the definitions for $p(\cdot)$, $P(\cdot,\cdot)$, $q(\cdot)$, $Q(\cdot,\cdot)$, $r(\cdot)$, and $s(\cdot)$ given in Section 5.1, define the following averages, which will be useful for the asymptotic linear representation of $\hat{\tau}^{\mathrm{CIC}}$:

$$\hat{\mu}^p=\frac{1}{N_{00}}\sum_{i=1}^{N_{00}}p(Y_{00,i}),\quad \hat{\mu}^P=\frac{1}{N_{00}}\frac{1}{N_{10}}\sum_{i=1}^{N_{00}}\sum_{j=1}^{N_{10}}P(Y_{00,i},Y_{10,j}),$$

$$\hat{\mu}^q=\frac{1}{N_{01}}\sum_{i=1}^{N_{01}}q(Y_{01,i}),\quad \hat{\mu}^Q=\frac{1}{N_{01}}\frac{1}{N_{10}}\sum_{i=1}^{N_{01}}\sum_{j=1}^{N_{10}}Q(Y_{01,i},Y_{10,j}),$$

$$\hat{\mu}^r=\frac{1}{N_{10}}\sum_{i=1}^{N_{10}}r(Y_{10,i}),\quad \hat{\mu}^s=\frac{1}{N_{11}}\sum_{i=1}^{N_{11}}s(Y_{11,i}).$$

LEMMA A.7: *Suppose Assumption* 5.1 *holds. Then*

$$\hat{\mu}^p-\hat{\mu}^P=o_p(N^{-1/2})\quad \textit{and}\quad \hat{\mu}^q-\hat{\mu}^Q=o_p(N^{-1/2}).$$

PROOF: Given $\hat{\mu}^P$ is a two-sample V-statistic, define $P_1(y)=\mathbb{E}[P(y,Y_{10})]$ and $P_2(y)=\mathbb{E}[P(Y_{00},y)]$. Standard theory for V-statistics implies that, under the smoothness and support conditions implied by Assumption 5.1,

$$\hat{\mu}^P=\frac{1}{N_{00}}\sum_{i=1}^{N_{00}}P_1(Y_{00,i})+\frac{1}{N_{10}}\sum_{i=1}^{N_{10}}P_2(Y_{10,i})+o_p(N^{-1/2}).$$

Because $P_1(y)=p(y)$ and $P_2(y)=0$, the result follows. The argument for $\hat{\mu}^Q$ is analogous. *Q.E.D.*

LEMMA A.8 —Consistency and Asymptotic Linearity: *Suppose Assumption* 5.1 *holds. Then*

$$\frac{1}{N_{10}}\sum_{i=1}^{N_{10}}\hat{F}_{Y,01}^{-1}(\hat{F}_{Y,00}(Y_{10,i}))\xrightarrow{p}\mathbb{E}\big[F_{Y,01}^{-1}(F_{Y,00}(Y_{10}))\big]$$

and

$$\frac{1}{N_{10}}\sum_{i=1}^{N_{10}}\hat{F}_{Y,01}^{-1}(\hat{F}_{Y,00}(Y_{10,i})) - \mathbb{E}\big[F_{Y,01}^{-1}(F_{Y,00}(Y_{10}))\big] - \hat{\mu}^p - \hat{\mu}^q - \hat{\mu}^r = o_p(N^{-1/2}).$$

PROOF: Because $\hat{F}_{Y,00}(z)$ converges to $F_{Y,00}(z)$ uniformly in z and because $\hat{F}_{Y,01}^{-1}(q)$ converges to $F_{Y,01}^{-1}(q)$ uniformly in q, it follows that $\hat{F}_{Y,01}^{-1}(\hat{F}_{Y,00}(z))$ converges to $F_{Y,01}^{-1}(F_{Y,00}(z))$ uniformly in z. Hence $(1/N_{10}) \times \sum_{i=1}^{N_{10}} \hat{F}_{Y,01}^{-1}(\hat{F}_{Y,00}(Y_{10,i}))$ converges to $(1/N_{10})\sum_{i=1}^{N_{10}} F_{Y,01}^{-1}(F_{Y,00}(Y_{10,i}))$, which by Assumption 5.1 and the law of large numbers converges to $\mathbb{E}[F_{Y,01}^{-1}(F_{Y,00}(Y_{10}))]$, which proves the first statement.

To prove the second statements, we will show that (A.5)–(A.7),

$$N^{1/2}\cdot\left(\frac{1}{N_{10}}\sum_{i=1}^{N_{10}}\hat{F}_{Y,01}^{-1}(\hat{F}_{Y,00}(Y_{10,i})) - \mathbb{E}\big[F_{Y,01}^{-1}(F_{Y,00}(Y_{10}))\big] - \hat{\mu}^p - \hat{\mu}^q - \hat{\mu}^r\right)$$

$$= N^{1/2}\cdot\left(\frac{1}{N_{10}}\sum_{i=1}^{N_{10}}\hat{F}_{Y,01}^{-1}(\hat{F}_{Y,00}(Y_{10,i})) - \frac{1}{N_{10}}\sum_{i=1}^{N_{10}}F_{Y,01}^{-1}(\hat{F}_{Y,00}(Y_{10,i})) - \hat{\mu}^q\right) \tag{A.5}$$

$$+ N^{1/2}\cdot\left(\frac{1}{N_{10}}\sum_{i=1}^{N_{10}}F_{Y,01}^{-1}(\hat{F}_{Y,00}(Y_{10,i})) - \frac{1}{N_{10}}\sum_{i=1}^{N_{10}}F_{Y,01}^{-1}(F_{Y,00}(Y_{10,i})) - \hat{\mu}^p\right) \tag{A.6}$$

$$+ N^{1/2}\cdot\left(\frac{1}{N_{10}}\sum_{i=1}^{N_{10}}F_{Y,01}^{-1}(F_{Y,00}(Y_{10,i})) - \mathbb{E}\big[F_{Y,01}^{-1}(F_{Y,00}(Y_{10}))\big] - \hat{\mu}^r\right), \tag{A.7}$$

are $o_p(1)$. First, (A.7) is equal to zero. Next, because $\hat{\mu}^p = \hat{\mu}^P + o_p(N^{-1/2})$ and $\hat{\mu}^q = \hat{\mu}^Q + o_p(N^{-1/2})$, it is sufficient to show that (A.5) and (A.6) with $\hat{\mu}^p$ and $\hat{\mu}^q$ replaced by $\hat{\mu}^Q$ and $\hat{\mu}^P$ are $o_p(1)$.

First, consider (A.5). By the TI,

$$
\begin{aligned}
&N^{1/2}\left|\frac{1}{N_{10}}\sum_{i=1}^{N_{10}}\hat{F}_{Y,01}^{-1}(\hat{F}_{Y,00}(Y_{10,i}))-\frac{1}{N_{10}}\sum_{i=1}^{N_{10}}F_{Y,01}^{-1}(\hat{F}_{Y,00}(Y_{10,i}))-\hat{\mu}^{Q}\right| \\
&\le N^{1/2}\left|\frac{1}{N_{10}}\sum_{i=1}^{N_{10}}\hat{F}_{Y,01}^{-1}(\hat{F}_{Y,00}(Y_{10,i}))-\frac{1}{N_{10}}\sum_{i=1}^{N_{10}}F_{Y,01}^{-1}(\hat{F}_{Y,00}(Y_{10,i}))\right. \qquad \text{(A.8)} \\
&\qquad +\frac{1}{N_{10}}\frac{1}{N_{01}}\sum_{i=1}^{N_{10}}\sum_{j=1}^{N_{01}}\frac{1}{f_{Y,01}(F_{Y,01}^{-1}(\hat{F}_{Y,00}(Y_{10,i})))} \\
&\qquad \left.\times\left(\mathbb{1}\{F_{Y,01}(Y_{01,j})\le\hat{F}_{Y,00}(Y_{10,i})\}-\hat{F}_{Y,00}(Y_{10,i})\right)\right| \\
&\quad +N^{1/2}\left|-\frac{1}{N_{10}}\frac{1}{N_{01}}\sum_{i=1}^{N_{10}}\sum_{j=1}^{N_{01}}\frac{1}{f_{Y,01}(F_{Y,01}^{-1}(\hat{F}_{Y,00}(Y_{10,i})))}\right. \qquad \text{(A.9)} \\
&\qquad \left.\times\left(\mathbb{1}\{F_{Y,01}(Y_{01,j})\le\hat{F}_{Y,00}(Y_{10,i})\}-\hat{F}_{Y,00}(Y_{10,i})\right)-\hat{\mu}^{Q}\right|.
\end{aligned}
$$

Equation (A.8) can be bounded by

$$
\begin{aligned}
&N^{1/2}\frac{1}{N_{10}}\sum_{i=1}^{N_{10}}\left|\hat{F}_{Y,01}^{-1}(\hat{F}_{Y,00}(Y_{10,i}))-F_{Y,01}^{-1}(\hat{F}_{Y,00}(Y_{10,i}))\right. \\
&\qquad +\frac{1}{N_{01}}\sum_{j=1}^{N_{01}}\frac{1}{f_{Y,01}(F_{Y,01}^{-1}(\hat{F}_{Y,00}(Y_{10,i})))} \\
&\qquad \left.\times\left(\mathbb{1}\{F_{Y,01}(Y_{01,j})\le\hat{F}_{Y,00}(Y_{10,i})\}-\hat{F}_{Y,00}(Y_{10,i})\right)\right| \\
&\le N^{1/2}\sup_{q}\left|\hat{F}_{Y,01}^{-1}(q)-F_{Y,01}^{-1}(q)\right. \\
&\qquad \left.+\frac{1}{N_{01}}\sum_{j=1}^{N_{01}}\frac{1}{f_{Y,01}(F_{Y,01}^{-1}(q))}\left(\mathbb{1}\{F_{Y,01}(Y_{01,j})\le q\}-q\right)\right| \\
&= N^{1/2}\sup_{q}\left|\hat{F}_{Y,01}^{-1}(q)-F_{Y,01}^{-1}(q)\right. \\
&\qquad \left.+\frac{1}{f_{Y,01}(F_{Y,01}^{-1}(q))}\left(\hat{F}_{Y,01}(F_{Y,01}^{-1}(q))-q\right)\right|,
\end{aligned}
$$

which is $o_p(1)$ by Lemma A.6. Next, consider (A.9):

$$
\begin{aligned}
N^{1/2}\Bigg|&\frac{1}{N_{10}}\frac{1}{N_{01}}\sum_{i=1}^{N_{10}}\sum_{j=1}^{N_{01}}\frac{1}{f_{Y,01}(F_{Y,01}^{-1}(\hat{F}_{Y,00}(Y_{10,i})))}\\
&\times\left(\mathbb{1}\{F_{Y,01}(Y_{01,j})\le\hat{F}_{Y,00}(Y_{10,i})\}-\hat{F}_{Y,00}(Y_{10,i})\right)\\
&-\frac{1}{N_{10}}\frac{1}{N_{01}}\sum_{i=1}^{N_{10}}\sum_{j=1}^{N_{01}}\frac{1}{f_{Y,01}(F_{Y,01}^{-1}(F_{Y,00}(Y_{10,i})))}\\
&\times\left(\mathbb{1}\{F_{Y,01}(Y_{01,j})\le F_{Y,00}(Y_{10,i})\}-F_{Y,00}(Y_{10,i})\right)\Bigg|\\
=N^{1/2}\Bigg|&\frac{1}{N_{10}}\sum_{i=1}^{N_{10}}\frac{1}{f_{Y,01}(F_{Y,01}^{-1}(\hat{F}_{Y,00}(Y_{10,i})))}\\
&\times\left(\hat{F}_{Y,01}\left(F_{Y,01}^{-1}(\hat{F}_{Y,00}(Y_{10,i}))\right)-\hat{F}_{Y,00}(Y_{10,i})\right)\\
&-\frac{1}{N_{10}}\sum_{i=1}^{N_{10}}\frac{1}{f_{Y,01}(F_{Y,01}^{-1}(F_{Y,00}(Y_{10,i})))}\\
&\times\left(\hat{F}_{Y,01}\left(F_{Y,01}^{-1}(F_{Y,00}(Y_{10,i}))\right)-F_{Y,00}(Y_{10,i})\right)\Bigg|.
\end{aligned}
$$

By the TI, this can be bounded by

$$
\begin{aligned}
&\text{(A.10)}\quad N^{1/2}\Bigg|\frac{1}{N_{10}}\sum_{i=1}^{N_{10}}\frac{1}{f_{Y,01}(F_{Y,01}^{-1}(\hat{F}_{Y,00}(Y_{10,i})))}\\
&\qquad\times\left(\hat{F}_{Y,01}\left(F_{Y,01}^{-1}(\hat{F}_{Y,00}(Y_{10,i}))\right)-\hat{F}_{Y,00}(Y_{10,i})\right)\\
&\qquad-\frac{1}{N_{10}}\sum_{i=1}^{N_{10}}\frac{1}{f_{Y,01}(F_{Y,01}^{-1}(\hat{F}_{Y,00}(Y_{10,i})))}\\
&\qquad\times\left(\hat{F}_{Y,01}\left(F_{Y,01}^{-1}(F_{Y,00}(Y_{10,i}))\right)-F_{Y,00}(Y_{10,i})\right)\Bigg|\\
&\text{(A.11)}\quad+N^{1/2}\Bigg|\frac{1}{N_{10}}\sum_{i=1}^{N_{10}}\frac{1}{f_{Y,01}(F_{Y,01}^{-1}(\hat{F}_{Y,00}(Y_{10,i})))}\\
&\qquad\times\left(\hat{F}_{Y,01}\left(F_{Y,01}^{-1}(F_{Y,00}(Y_{10,i}))\right)-F_{Y,00}(Y_{10,i})\right)
\end{aligned}
$$

$$- \frac{1}{N_{10}} \sum_{i=1}^{N_{10}} \frac{1}{f_{Y,01}(F_{Y,01}^{-1}(F_{Y,00}(Y_{10,i})))} \times \left(\hat{F}_{Y,01}\left(F_{Y,01}^{-1}(F_{Y,00}(Y_{10,i}))\right) - F_{Y,00}(Y_{10,i})\right)\Bigg|.$$

Equation (A.10) can be bounded by

$$\begin{aligned} N^{1/2} \sup_q \left| \frac{1}{f_{Y,01}(F_{Y,01}^{-1}(q))} \right| \cdot \sup_y \Big| & \hat{F}_{Y,01}\left(F_{Y,01}^{-1}(\hat{F}_{Y,00}(y))\right) - \hat{F}_{Y,00}(y) \\ & - \left(\hat{F}_{Y,01}\left(F_{Y,01}^{-1}(F_{Y,00}(y))\right) - F_{Y,00}(y)\right)\Big| \\ \leq N^{1/2} \cdot C \cdot \sup_y \Big| & \hat{F}_{Y,01}\left(F_{Y,01}^{-1}(\hat{F}_{Y,00}(y))\right) - \hat{F}_{Y,01}\left(F_{Y,01}^{-1}(F_{Y,00}(y))\right) \\ & - \left(F_{Y,01}\left(F_{Y,01}^{-1}(\hat{F}_{Y,00}(y))\right) - F_{Y,00}(y)\right)\Big|. \end{aligned}$$

To see that this is $o_p(1)$, we apply Lemma A.5. Take $\delta = 1/3$ and $\eta = 1/2$. Then $\hat{F}_{Y,00}(y) - F_{Y,00}(y) = o_p(N^{-\delta})$, and thus the conditions for Lemma A.5 are satisfied and so (A.11) is $o_p(1)$. Equation (A.11) can be bounded by

$$\begin{aligned} N^{1/4} \sup_q & \left| \frac{1}{f_{Y,01}(F_{Y,01}^{-1}(\hat{F}_{Y,00}(Y_{10,i})))} - \frac{1}{f_{Y,01}(F_{Y,01}^{-1}(F_{Y,00}(Y_{10,i})))} \right| \\ & \times N^{1/4} \sup_q \left| \hat{F}_{Y,01}\left(F_{Y,01}^{-1}(F_{Y,00}(Y_{10,i}))\right) - F_{Y,00}(Y_{10,i}) \right|. \end{aligned}$$

Both factors are $o_p(1)$, so (A.11) is $o_p(1)$.

Second, consider (A.6):

$$\begin{aligned} N^{1/2} & \left| \frac{1}{N_{10}} \sum_{i=1}^{N_{10}} F_{Y,01}^{-1}(\hat{F}_{Y,00}(Y_{10,i})) - \frac{1}{N_{10}} \sum_{i=1}^{N_{10}} F_{Y,01}^{-1}(F_{Y,00}(Y_{10,i})) - \hat{\mu}^P \right| \\ = N^{1/2} & \left| \frac{1}{N_{10}} \sum_{i=1}^{N_{10}} \Bigg(F_{Y,01}^{-1}(\hat{F}_{Y,00}(y)) - F_{Y,01}^{-1}(F_{Y,00}(Y_{10,i})) \right. \\ & \qquad - \frac{1}{f_{Y,01}(F_{Y,01}^{-1}(F_{Y,00}(Y_{10,i})))} \\ & \qquad \left. \times \frac{1}{N_{00}} \sum_{j=1}^{N_{00}} \left(\mathbb{1}\{Y_{00,j} < Y_{10,i}\} - F_{Y,00}(Y_{10,i})\right) \Bigg) \right| \end{aligned}$$

$$\leq N^{1/2} \sup_y \Bigg| F_{Y,01}^{-1}(\hat{F}_{Y,00}(y)) - F_{Y,01}^{-1}(F_{Y,00}(y)) - \frac{1}{f_{Y,01}(F_{Y,01}^{-1}(F_{Y,00}(y)))} \frac{1}{N_{00}} \sum_{i=1}^{N_{00}} \big(\mathbb{1}\{Y_{00,i} < y\} - F_{Y,00}(y) \big) \Bigg|$$

$$= N^{1/2} \sup_y \Bigg| F_{Y,01}^{-1}(\hat{F}_{Y,00}(y)) - F_{Y,01}^{-1}(F_{Y,00}(y)) - \frac{1}{f_{Y,01}(F_{Y,01}^{-1}(F_{Y,00}(y)))} (\hat{F}_{Y,00}(y) - F_{Y,00}(y)) \Bigg|.$$

Expanding $F_{Y,01}^{-1}(\hat{F}_{Y,00}(y))$ around $F_{Y,00}(y)$ implies that this can be bounded by

$$N^{1/2} \sup_y \left| \frac{1}{f_{Y,01}(y)^3} \frac{\partial f_{Y,01}}{\partial y}(y) \right| \cdot \sup_y |\hat{F}_{Y,00}(y) - F_{Y,00}(y)|^2,$$

which is $o_p(1)$ by Lemma A.2.

Finally, the third term (A.7) is equal to zero. *Q.E.D.*

LEMMA A.9—Asymptotic Normality: *Suppose Assumption* 5.1 *holds*. *Then*

$$\sqrt{N} \left(\frac{1}{N_{10}} \sum_{i=1}^{N_{10}} \hat{F}_{Y,01}^{-1}(\hat{F}_{Y,00}(Y_{10,i})) - \mathbb{E}\big[F_{Y,01}^{-1}(F_{Y,00}(Y_{10})) \big] \right) \xrightarrow{d} \mathcal{N}\left(0, \frac{V^p}{\alpha_{00}} + \frac{V^q}{\alpha_{01}} + \frac{V^r}{\alpha_{10}} \right).$$

PROOF: Because of Lemma A.8, it is sufficient to show that

$$\sqrt{N}(\hat{\mu}^p + \hat{\mu}^q + \hat{\mu}^r) \xrightarrow{d} \mathcal{N}(0, V^p/\alpha_{00} + V^q/\alpha_{01} + V^r/\alpha_{10}).$$

Conditional on N_{gt}, all three components $\hat{\mu}^p$, $\hat{\mu}^q$, and $\hat{\mu}^r$ are sample averages of independent and identically distributed random variables. Given the assumptions on the distributions of Y_{gt}, all the moments of these functions exist, and, therefore, central limit theorems apply and the result follows directly. *Q.E.D.*

PROOF OF THEOREM 5.1: Apply Lemmas A.8 and A.9, which give us the asymptotic distribution of $\sum \hat{F}_{Y,01}^{-1}(\hat{F}_{Y,00}(Y_{10i}))/N_{10}$. We are interested in the large sample behavior of $\sum Y_{11i}/N_{11} - \sum \hat{F}_{Y,01}^{-1}(\hat{F}_{Y,00}(Y_{10i}))/N_{10}$. Whereas $\sum_i Y_{11i}/N_{11} = \hat{\mu}^s$ is asymptotically independent of $\sum \hat{F}_{Y,01}^{-1}(\hat{F}_{Y,00}(Y_{10i}))/N_{10}$, this just leads to the extra variance term V_{11}/α_{11}. *Q.E.D.*

Before proving Theorem 5.2, we state two preliminary lemmas. Proofs are provided in the supplement.

LEMMA A.10: *Suppose that for* $h_1, \hat{h}_1 : \mathbb{Y}_1 \to \mathbb{R}$, *and* $h_2, \hat{h}_2 : \mathbb{Y}_2 \to \mathbb{R}$, $\sup_{y \in \mathbb{Y}_1} |\hat{h}_1(y) - h_1(y)| \xrightarrow{p} 0$, $\sup_{y \in \mathbb{Y}_2} |\hat{h}_2(y) - h_2(y)| \xrightarrow{p} 0$, $\sup_{y \in \mathbb{Y}_1} |h_1(y)| < \overline{h_1} < \infty$, *and* $\sup_{y \in \mathbb{Y}_2} |h_2(y)| < \overline{h_2} < \infty$. *Then*

$$\sup_{y_1 \in \mathbb{Y}_1, y_2 \in \mathbb{Y}_2} |\hat{h}_1(y_1)\hat{h}_2(y_2) - h_1(y_1)h_2(y_2)| \longrightarrow 0.$$

LEMMA A.11: *Suppose that for* $h_1, \hat{h}_1 : \mathbb{Y}_1 \to \mathbb{Y}_2 \subset \mathbb{R}$ *and* $h_2 : \mathbb{Y}_2 \to \mathbb{R}$, $\sup_{y \in \mathbb{Y}_1} |\hat{h}_1(y) - h_1(y)| \xrightarrow{p} 0$ *and* $\sup_{y \in \mathbb{Y}_2} |\hat{h}_2(y) - h_2(y)| \xrightarrow{p} 0$, *and suppose that* $h_2(y)$ *is continuously differentiable with its derivative bounded in absolute value by* $\overline{h_2'} < \infty$. *Then*

$$\text{(A.12)} \quad \sup_{y \in \mathbb{Y}_1} \left| \hat{h}_2(\hat{h}_1(y)) - h_2(h_1(y)) \right| \xrightarrow{p} 0.$$

PROOF OF THEOREM 5.2: Let $\underline{f} = \inf_{y,g,t} f_{Y,gt}(y)$, $\overline{f} = \sup_{y,g,t} f_{Y,gt}(y)$, and $\overline{f'} = \sup_{y,g,t} (\partial f_{Y,gt}/\partial y)(y)$. Also let $C_p = \sup_{y_{00}, y_{10}} p(y_{00}, y_{10})$, $C_q = \sup_{y_{01}, y_{10}} q(y_{01}, y_{10})$, and $C_r = \sup_{y_{10}} r(y_{10})$. By Assumption 5.1, $\underline{f} > 0$, $\overline{f} < \infty$, $\overline{f'} < \infty$, and $C_p, C_q, C_r < \infty$.

It suffices to show $\hat{\alpha}_{gt} \xrightarrow{p} \alpha_{gt}$ for all $g, t = 0, 1$, and $\hat{V}^p \xrightarrow{p} V^p$, $\hat{V}^q \xrightarrow{p} V^q$, $\hat{V}^r \xrightarrow{p} V^r$, and $\hat{V}^s \xrightarrow{p} V^s$. Consistency of $\hat{\alpha}_{gt}$ and $\hat{V}^s$ is immediate. Next consider consistency of $\hat{V}^p$. The proof is broken up into three steps: the first step is to prove uniform consistency of $\hat{\hat{f}}_{Y,00}(y)$, the second step is to prove uniform consistency of $\hat{P}(y_{00}, y_{10})$ in both its arguments, and the third step is to prove consistency of $\hat{V}^p$ given uniform consistency of $\hat{P}(y_{00}, y_{10})$.

For uniform consistency of $\hat{\hat{f}}_{Y,00}(y)$, first note that, for all $0 < \delta < 1/2$, we have, by Lemmas A.2 and A.3,

$$\sup_{y \in \mathbb{Y}_{gt}} N_{gt}^{\delta} \cdot |\hat{F}_{Y,gt}(y) - F_{Y,gt}(y)| \xrightarrow{p} 0 \quad \text{and}$$

$$\sup_{q \in [0,1]} N_{gt}^{\delta} \cdot |\hat{F}_{Y,gt}^{-1}(q) - F_{Y,gt}^{-1}(q)| \xrightarrow{p} 0.$$

Now consider first the case with $y < \tilde{Y}_{gt}$:

$$\sup_{y < \tilde{Y}_{gt}} |\hat{\hat{f}}_{Y,gt}(y) - f_{Y,gt}(y)|$$

$$= \sup_{y < \tilde{Y}_{gt}} \left| \frac{\hat{F}_{Y,gt}(y + N^{-1/3}) - \hat{F}_{Y,gt}(y)}{N^{-1/3}} - f_{Y,gt}(y) \right|$$

$$
\begin{aligned}
&\leq \sup_{y<\tilde{Y}_{gt}} \left| \frac{\hat{F}_{Y,gt}(y+N^{-1/3}) - \hat{F}_{Y,gt}(y)}{N^{-1/3}} - \frac{F_{Y,gt}(y+N^{-1/3}) - F_{Y,gt}(y)}{N^{-1/3}} \right| \\
&\quad + \sup_{y<\tilde{Y}_{gt}} \left| \frac{F_{Y,gt}(y+N^{-1/3}) - F_{Y,gt}(y)}{N^{-1/3}} - f_{Y,gt}(y) \right| \\
&\leq \sup_{y<\tilde{Y}_{gt}} \left| \frac{\hat{F}_{Y,gt}(y+N^{-1/3}) - F_{Y,gt}(y+N^{-1/3})}{N^{-1/3}} - \frac{\hat{F}_{Y,gt}(y) - F_{Y,gt}(y)}{N^{-1/3}} \right| \\
&\quad + N^{-1/3} \left| \frac{\partial f_{Y,gt}}{\partial y}(\tilde{y}) \right| \\
&\leq 2N^{1/3} \sup_{y\in\mathbb{Y}_{gt}} |\hat{F}_{Y,gt}(y) - F_{Y,gt}(y)| + N^{-1/3} \sup_{y\in\mathbb{Y}_{gt}} \left| \frac{\partial f_{Y,gt}}{\partial y}(y) \right| \\
&\xrightarrow{p} 0,
\end{aligned}
$$

where $\tilde{y}$ is some value in the support $\mathbb{Y}_{gt}$. The same argument shows that $\sup_{y\geq\tilde{Y}_{gt}} |\hat{f}_{Y,gt}(y) - f_{Y,gt}(y)| \xrightarrow{p} 0$, which, combined with the earlier part, shows that $\sup_{y\in\mathbb{Y}_{gt}} |\hat{f}_{Y,gt}(y) - f_{Y,gt}(y)| \xrightarrow{p} 0$.

The second step is to show uniform consistency of $\hat{P}(y_{00}, y_{10})$. By boundedness of the derivative of $F^{-1}_{Y,01}(q)$, and uniform convergence of $\hat{F}^{-1}_{Y,01}(q)$ and $\hat{F}_{Y,00}(y)$, Lemma A.11 implies uniform convergence of $\hat{F}^{-1}_{Y,01}(\hat{F}_{Y,00}(y))$ to $F^{-1}_{Y,01}(F_{Y,00}(y))$. This in turn, combined with uniform convergence of $\hat{f}_{Y,01}(y)$ and another application of Lemma A.11, implies uniform convergence of $\hat{f}_{Y,01}(\hat{F}^{-1}_{Y,01}(\hat{F}_{Y,00}(y_{10})))$ to $f_{Y,01}(F^{-1}_{Y,01}(F_{Y,00}(y_{10})))$. Because $f_{Y,01}(y)$ is bounded away from zero, this implies uniform convergence of $1/\hat{f}_{Y,01}(\hat{F}^{-1}_{Y,01}(\hat{F}_{Y,00}(y_{10})))$ to $1/f_{Y,01}(F^{-1}_{Y,01}(F_{Y,00}(y_{10})))$. Finally, using Lemma A.10 then gives uniform convergence of $\hat{P}(y_{00}, y_{10})$ to $P(y_{00}, y_{10})$, completing the second step of the proof.

The third step is to show consistency of $\hat{V}^p$ given uniform convergence of $\hat{P}(y_{00}, y_{10})$. For any $\varepsilon > 0$, let $\eta = \min(\sqrt{\varepsilon/2}, \varepsilon/(4C_p))$ (where, as defined before, $C_p = \sup_{y,z} P(y, z)$). Then for N large enough so that $\sup_{y_{00},y_{10}} |\hat{P}(y_{00}, y_{10}) - P(y_{00}, y_{10})| < \eta$, it follows that

$$
\begin{aligned}
&\sup_{y_{00}} \left| \frac{1}{N_{10}} \sum_{j=1}^{N_{10}} \hat{P}(y_{00}, Y_{10,j}) - \frac{1}{N_{10}} \sum_{j=1}^{N_{10}} P(y_{00}, Y_{10,j}) \right| \\
&\quad \leq \sup_{y_{00}} \frac{1}{N_{10}} \sum_{j=1}^{N_{10}} |\hat{P}(y_{00}, Y_{10,j}) - P(y_{00}, Y_{10,j})| < \eta
\end{aligned}
$$

and thus, using $A^2 - B^2 = (A-B)^2 + 2B(A-B)$,

$$\sup_{y_{00}} \left| \left[\frac{1}{N_{10}} \sum_{j=1}^{N_{10}} \hat{P}(y_{00}, Y_{10,j}) \right]^2 - \left[\frac{1}{N_{10}} \sum_{j=1}^{N_{10}} P(y_{00}, Y_{10,j}) \right]^2 \right| < \eta^2 + 2C_p \eta \leq \varepsilon.$$

Hence

$$\left| \frac{1}{N_{00}} \sum_{i=1}^{N_{00}} \left[\frac{1}{N_{10}} \sum_{j=1}^{N_{10}} \hat{P}(Y_{00,i}, Y_{10,j}) \right]^2 - \frac{1}{N_{00}} \sum_{i=1}^{N_{00}} \left[\frac{1}{N_{10}} \sum_{j=1}^{N_{10}} P(Y_{00,i}, Y_{10,j}) \right]^2 \right| \leq \varepsilon.$$

Thus it remains to prove that

$$V^p - \frac{1}{N_{00}} \sum_{i=1}^{N_{00}} \left[\frac{1}{N_{10}} \sum_{j=1}^{N_{10}} P(Y_{00,i}, Y_{10,j}) \right]^2 \xrightarrow{p} 0.$$

By boundedness of $P(y_{00}, y_{10})$, it follows that $(1/N_{10}) \sum_{j=1}^{N_{10}} P(y, Y_{10,j}) - \mathbb{E}[P(y, Y_{10})] = (1/N_{10}) \sum_{j=1}^{N_{10}} P(y, Y_{10,j}) - p(y) \xrightarrow{p} 0$ uniformly in y. Hence,

$$\frac{1}{N_{00}} \sum_{i=1}^{N_{00}} \left[\frac{1}{N_{10}} \sum_{j=1}^{N_{10}} P(Y_{00,i}, Y_{10,j}) \right]^2 - \frac{1}{N_{00}} \sum_{i=1}^{N_{00}} p(Y_{00,i})^2 \xrightarrow{p} 0.$$

Finally, by the law of large numbers, $\sum_{i=1}^{N_{00}} p(Y_{00,i})^2 / N_{00} - V^p \xrightarrow{p} 0$, implying consistency of $\hat{V}^p$. Consistency of $\hat{V}^q$ and $\hat{V}^r$ follows the same pattern of first establishing uniform consistency of $\hat{Q}(y_{01}, y_{10})$ and $\hat{r}(y)$, respectively, followed by using the law of large numbers. The proofs are therefore omitted. *Q.E.D.*

Next we establish an alternative representation of the bounds on the distribution function, as well as an analytic representation of bounds on the average treatment effect.

LEMMA A.12—Bounds on the Average Treatment Effect: *Suppose Assumptions* 3.1, 3.3, 3.4, 4.2, 4.3, *and* 5.2 *hold*. *Suppose that the support of* $\mathbb{Y}$ *is a finite set*. *Then*:

(i) $F^{\mathrm{LB}}_{Y^N,11}(y) = \Pr(\overline{k}(Y_{10}) \leq y)$ *and* $F^{\mathrm{UB}}_{Y^N,11}(y) = \Pr(\underline{k}(Y_{10}) \leq y)$.

(ii) *The average treatment effect, τ, satisfies*

$$\tau \in \left[\mathbb{E}[Y_{11}^I] - \mathbb{E}\left[F_{Y,01}^{-1}(F_{Y,00}(Y_{10}))\right], \mathbb{E}[Y_{11}^I] - \mathbb{E}\left[F_{Y,01}^{-1}(\underline{F}_{Y,00}(Y_{10}))\right]\right].$$

PROOF: Let $\mathbb{Y}_{00} = \{\lambda_1, \ldots, \lambda_L\}$ and $\mathbb{Y}_{01} = \{\gamma_1, \ldots, \gamma_M\}$ be the supports of Y_{00} and Y_{01}, respectively.[40] By Assumption 3.4 the supports of Y_{10} and Y_{11}^N are subsets of these.

Fix y. Let $l(y) = \max\{l = 1, \ldots, L : \underline{k}(\lambda_l) \leq y\}$. Consider two cases: (i) $l(y) < L$ and (ii) $l(y) = L$. Start with case (i). Then, $\underline{k}(\lambda_{l(y)+1}) > y$. Also, since $\underline{k}(y)$ is nondecreasing in y,

$$\tilde{F}_{Y^N,11}^{\mathrm{UB}}(y) \equiv \Pr(\underline{k}(Y_{10}) \leq y) = \Pr(Y_{10} \leq \lambda_{l(y)}) = F_{Y,10}(\lambda_{l(y)}).$$

Define $\gamma(y) \equiv \underline{k}(\lambda_{l(y)})$ and $\gamma'(y) \equiv \underline{k}(\lambda_{l(y)+1})$ so that $\gamma(y) \leq y < \gamma'(y)$. Also define for $j \in \{1, \ldots, L\}$, $q_j = F_{Y_{00}}(\lambda_j)$ and note that by definition of $\underline{F}_{Y,00}$, $\underline{F}_{Y,00}(\lambda_j) = q_{j-1}$. Define $p(y) \equiv F_{Y,01}(y)$. Because $y \geq \underline{k}(\lambda_{l(y)}) = F_{Y,01}^{-1}(\underline{F}_{Y,00}(\lambda_{l(y)}))$ (the inequality follows from the definition of $l(y)$; the equality follows from the definition of $\underline{k}(y)$), applying the nondecreasing function $F_{Y,01}(\cdot)$ to both sides of the inequality yields $p(y) = F_{Y,01}(y) \geq F_{Y,01}(F_{Y,01}^{-1}(\underline{F}_{Y,00}(\lambda_{l(y)})))$. By the definition of the inverse distribution function, $F_Y(F_Y^{-1}(q)) \geq q$, so that $p(y) \geq \underline{F}_{Y,00}(\lambda_{l(y)}) = q_{l(y)-1}$. Because $l(y) < L$, Assumption 5.2 rules out equality of $F_{Y,01}(\gamma_m)$ and $F_{Y,00}(\lambda_j)$ and, therefore $p(y) > q_{l(y)-1}$. Also, $F_{Y,01}^{-1}(p(y)) = F_{Y,01}^{-1}(F_{Y,01}(y)) \leq y < \gamma'(y)$ and, substituting in definitions, $\gamma'(y) = F_{Y,01}^{-1}(\underline{F}_{Y,00}(\lambda_{l(y)+1})) = F_{Y,01}^{-1}(q_{l(y)})$. Putting the latter two conclusions together, we conclude that $F_{Y,01}^{-1}(p(y)) < F_{Y,01}^{-1}(q_{l(y)})$, which implies $p(y) < q_{l(y)}$. Whereas we have now established $q_{l(y)-1} < p(y) < q_{l(y)}$, it follows by the definition of the inverse function that $F_{Y,00}^{-1}(p(y)) = \lambda_{l(y)}$. Hence,

$$\begin{aligned} F_{Y_{11}^N}^{\mathrm{UB}}(y) &= F_{Y,10}\left(F_{Y,00}^{-1}(F_{Y,01}(y))\right) \\ &= F_{Y,10}\left(F_{Y,00}^{-1}(p(y))\right) = F_{Y,10}(\lambda_{l(y)}) = \tilde{F}_{Y_{11}^N}^{\mathrm{UB}}(y). \end{aligned}$$

This proves the first part of the lemma for the upper bound for case (i).

In case (ii), $\underline{k}(\lambda_L) \leq y$, implying that $\tilde{F}_{Y^N,11}^{\mathrm{UB}}(y) \equiv \Pr(\underline{k}(Y_{10}) \leq y) = \Pr(Y_{10} \leq \lambda_L) = 1$. Applying the same argument as before, one can show that $p(y) \equiv F_{Y,01}(y) \geq \underline{F}_{Y,00}(\lambda_L)$, implying $F_{Y,00}^{-1}(p(y)) = \lambda_L$ and, hence, $F_{Y^N,11}^{\mathrm{UB}}(y) = F_{Y,10}(\lambda_L) = 1 = \tilde{F}_{Y^N,11}^{\mathrm{UB}}(y)$.

The result for the lower bound follows the same pattern and is omitted here. The second part of the lemma follows because we have established that $\underline{k}(Y_{10})$ has distribution $F_{Y^N,11}^{\mathrm{UB}}(\cdot)$ and $\overline{k}(Y_{10})$ has distribution $F_{Y^N,11}^{\mathrm{LB}}(\cdot)$. *Q.E.D.*

[40]These supports can be the same.

Before proving Theorem 5.4 we need a preliminary result.

LEMMA A.13: *For all* $l = 1, \ldots, L$, $\sqrt{N}(\hat{\underline{k}}(\lambda_l) - \underline{k}(\lambda_l)) \xrightarrow{p} 0$ and $\sqrt{N} \times (\hat{\overline{k}}(\lambda_l) - \overline{k}(\lambda_l)) \xrightarrow{p} 0$.

PROOF: Define $\nu = \min_{l,m:\min(l,m)<L} |F_{00}(\lambda_l) - F_{01}(\lambda_m)|$. By Assumption 5.2 and the finite support assumption, $\nu > 0$. By uniform convergence of the empirical distribution function, there is for all $\varepsilon > 0$ an $N_{\varepsilon,\nu}$ such that for $N \geq N_{\varepsilon,\nu}$ we have

$$\Pr\Big(\sup_y |\hat{F}_{Y,00}(y) - F_{Y,00}(y)| > \nu/3\Big) < \varepsilon/4,$$

$$\Pr\Big(\sup_y |\hat{F}_{Y,01}(y) - F_{Y,01}(y)| > \nu/3\Big) < \varepsilon/4$$

and

$$\Pr\Big(\sup_y |\hat{\underline{F}}_{Y,00}(y) - \underline{F}_{Y,00}(y)| > \nu/3\Big) < \varepsilon/4,$$

$$\Pr\Big(\sup_y |\hat{\underline{F}}_{Y,01}(y) - \underline{F}_{Y,01}(y)| > \nu/3\Big) < \varepsilon/4.$$

Now consider the case where

$$\text{(A.13)} \quad \sup_y |\hat{F}_{Y,00}(y) - F_{Y,00}(y)| \leq \nu/3,$$

$$\sup_y |\hat{F}_{Y,01}(y) - F_{Y,01}(y)| \leq \nu/3,$$

$$\sup_y |\hat{\underline{F}}_{Y,00}(y) - \underline{F}_{Y,00}(y)| \leq \nu/3, \quad \text{and}$$

$$\sup_y |\hat{\underline{F}}_{Y,01}(y) - \underline{F}_{Y,01}(y)| \leq \nu/3.$$

By the above argument the probability of (A.13) is larger than $1 - \varepsilon$ for $N \geq N_{\varepsilon,\nu}$. Hence, it can be made arbitrarily close to 1 by choosing N large enough.

Let $\lambda_m = F^{-1}_{Y,01}(q_{00,l})$. By Assumption 5.2, it follows that $F_{Y,01}(\lambda_{m-1}) < q_{00,l} = F_{Y,00}(\lambda_l) < F_{Y,01}(\lambda_m)$, with $F_{Y,01}(\lambda_m) - q_{00,l} > \nu$ and $q_{00,l} - F_{Y,01}(\lambda_{m-1}) > \nu$ by the definition of ν. Conditional on (A.13), we therefore have $\hat{F}_{Y,01}(\lambda_{m-1}) < \hat{F}_{Y,00}(\lambda_l) < \hat{F}_{Y,01}(\lambda_m)$. This implies $\hat{F}^{-1}_{Y,01}(\hat{F}_{Y,00}(\lambda_l)) = \lambda_m = F^{-1}_{Y,01}(F_{Y,00}(\lambda_l))$, and thus $\hat{\underline{k}}(\lambda_l) = \underline{k}(\lambda_l)$. Hence, for any $\eta, \varepsilon > 0$, for $N > N_{\varepsilon,\nu}$, we have

$$\begin{aligned}\Pr\big(\big|\sqrt{N}(\hat{\underline{k}}(\lambda_l) - \underline{k}(\lambda_l))\big| > \eta\big) &\leq 1 - \Pr\big(\big|\sqrt{N}(\hat{\underline{k}}(\lambda_l) - \underline{k}(\lambda_l))\big| = 0\big) \\ &\leq 1 - (1 - \varepsilon) = \varepsilon,\end{aligned}$$

which can be choosen arbitrarily small. The same argument applies to $\sqrt{N}(\hat{\overline{k}}(\lambda_l) - \overline{k}(\lambda_l))$, so it is therefore omitted. *Q.E.D.*

PROOF OF THEOREM 5.4: We prove only the first assertion; the second follows the same argument. Consider

$$\begin{aligned}\sqrt{N}(\hat{\tau}_{\mathrm{UB}} - \tau_{\mathrm{UB}}) &= \frac{1}{\sqrt{\alpha_{11}N_{11}}} \cdot \sum_{i=1}^{N_{11}} (Y_{11,i} - \mathbb{E}[Y_{11}]) \\ &\quad - \frac{1}{\sqrt{\alpha_{10}N_{10}}} \cdot \sum_{i=1}^{N_{10}} \big(\hat{\underline{k}}(Y_{10,i}) - \mathbb{E}[\underline{k}(Y_{10})]\big) \\ &= \frac{1}{\sqrt{\alpha_{11}N_{11}}} \cdot \sum_{i=1}^{N_{11}} (Y_{11,i} - \mathbb{E}[Y_{11}]) \\ &\quad - \frac{1}{\sqrt{\alpha_{10}N_{10}}} \cdot \sum_{i=1}^{N_{10}} \big(\underline{k}(Y_{10,i}) - \mathbb{E}[\underline{k}(Y_{10})]\big) \\ &\quad + \frac{1}{\sqrt{\alpha_{10}N_{10}}} \cdot \sum_{i=1}^{N_{10}} (\hat{\underline{k}}(Y_{10,i}) - \underline{k}(Y_{10})).\end{aligned}$$

By the central limit theorem, and independence of $\bar{Y}_{11}$ and $\underline{k}(\bar{Y}_{10})$, we have

$$\begin{aligned}&\frac{1}{\sqrt{\alpha_{11}N_{11}}} \cdot \sum_{i=1}^{N_{11}} (Y_{11,i} - \mathbb{E}[Y_{11}]) - \frac{1}{\sqrt{\alpha_{10}N_{10}}} \cdot \sum_{i=1}^{N_{10}} \big(\underline{k}(Y_{10,i}) - \mathbb{E}[\underline{k}(Y_{10})]\big) \\ &\quad \xrightarrow{d} \mathcal{N}\left(0, \frac{V^s}{\alpha_{11}} + \frac{V^r}{\alpha_{10}}\right).\end{aligned}$$

Hence all we need to prove is that $(1/\sqrt{\alpha_{10}N_{10}}) \cdot \sum_{i=1}^{N_{10}} (\hat{\underline{k}}(Y_{10,i}) - \underline{k}(Y_{10})) \xrightarrow{p} 0$. This expression can be bounded in absolute value by $\sqrt{N} \cdot \max_{l=1,\ldots,L} |\hat{\underline{k}}(\lambda_l) - \underline{k}(\lambda_l)|$. Because $\sqrt{N} \cdot |\hat{\underline{k}}(\lambda_l) - \underline{k}(\lambda_l)|$ converges to zero for each l by Lemma A.13, this converges to zero. *Q.E.D.*

PROOF OF THEOREM 6.2: The result in Corollary 6.1 implies that it is sufficient to show that $\sqrt{N}(\tilde{\kappa}_{\mathcal{J}} - \kappa_{\mathcal{J}}) \xrightarrow{d} \mathcal{N}(0, V_{\mathcal{J}})$. To show joint normality, we need to show that any arbitrary linear combinations of terms of the form the $\sqrt{N} \cdot (\tilde{\kappa}_{g_0,g_1,t_0,t_1} - \kappa_{g_0,g_1,t_0,t_1})$ are normally distributed. This follows from the asymptotic normality and independence of the $\hat{\mu}^p_{g,t}$, $\hat{\mu}^q_{g,t}$, $\hat{\mu}^r_{g,t}$, and $\hat{\mu}^s_{g,t}$, combined with their independence across groups and time periods. *Q.E.D.*

APPENDIX B: COVARIANCES OF $\sqrt{N}\hat{\kappa}_{g_0,g_1,t_0,t_1}$

Here we list, for all combinations of (g_0, g_1, t_0, t_1) and (g_0', g_1', t_0', t_1'), the covariance of $\sqrt{N}\hat{\kappa}_{g_0,g_1,t_0,t_1}$ and $\sqrt{N}\hat{\kappa}_{g_0',g_1',t_0',t_1'}$. Note that $t_1 > t_0$ and $t_1' > t_0'$. To avoid duplication, we also consider only the cases with $g_1 > g_0$ and $g_1' > g_0'$.

1. $g_0 = g_0'$, $g_1 = g_1'$, $t_0 = t_0'$, and $t_1 = t_1'$: $C = N \cdot \mathbb{E}[(\hat{\mu}^p_{g_0,g_1,t_0,t_1})^2] + N \cdot \mathbb{E}[(\hat{\mu}^q_{g_0,g_1,t_0,t_1})^2] + N \cdot \mathbb{E}[(\hat{\mu}^r_{g_0,g_1,t_0,t_1})^2] + N \cdot \mathbb{E}[(\hat{\mu}^s_{g_0,g_1,t_0,t_1})^2]$.

2. $g_0 = g_0'$, $g_1 = g_1'$, $t_0 = t_0'$, and $t_1 \neq t_1'$: $C = N \cdot \mathbb{E}[\hat{\mu}^p_{g_0,g_1,t_0,t_1}(Y_{g_0,t_0}) \cdot \hat{\mu}^p_{g_0,g_1,t_0,t_1'}(Y_{g_0,t_0})]/\alpha_{g_0,t_0} + N \cdot \mathbb{E}[\hat{\mu}^r_{g_0,g_1,t_0,t_1} \cdot \hat{\mu}^r_{g_0,g_1,t_0,t_1'}]$.

3. $g_0 = g_0'$, $g_1 = g_1'$, $t_0 \neq t_0'$, and $t_1 = t_1'$: $C = N \cdot \mathbb{E}[\hat{\mu}^q_{g_0,g_1,t_0,t_1} \cdot \hat{\mu}^q_{g_0,g_1,t_0',t_1}] + N \cdot \mathbb{E}[\hat{\mu}^s_{g_0,g_1,t_0,t_1} \cdot \hat{\mu}^s_{g_0,g_1,t_0',t_1}]$.

4. $g_0 = g_0'$, $g_1 = g_1'$, $t_0 \neq t_0'$, $t_1 \neq t_1'$, and $t_0' = t_1$: $C = N \cdot \mathbb{E}[\hat{\mu}^q_{g_0,g_1,t_0,t_1} \cdot \hat{\mu}^p_{g_0,g_1,t_1,t_1'}] + N \cdot \mathbb{E}[\hat{\mu}^s_{g_0,g_1,t_0,t_1} \cdot \hat{\mu}^r_{g_0,g_1,t_1,t_1'}]$.

5. $g_0 = g_0'$, $g_1 = g_1'$, $t_0 \neq t_0'$, $t_1 \neq t_1'$, and $t_0 = t_1'$: $C = N \cdot \mathbb{E}[\hat{\mu}^p_{g_0,g_1,t_0,t_1} \cdot \hat{\mu}^q_{g_0,g_1,t_0',t_0}] + N \cdot \mathbb{E}[\hat{\mu}^r_{g_0,g_1,t_0,t_1} \cdot \hat{\mu}^s_{g_0,g_1,t_0',t_0}]$.

6. $g_0 = g_0'$, $g_1 \neq g_1'$, $t_0 = t_0'$, and $t_1 = t_1'$: $C = N \cdot \mathbb{E}[\hat{\mu}^p_{g_0,g_1,t_0,t_1} \cdot \hat{\mu}^p_{g_0,g_1',t_0,t_1}] + N \cdot \mathbb{E}[\hat{\mu}^q_{g_0,g_1,t_0,t_1} \cdot \hat{\mu}^q_{g_0,g_1',t_0,t_1}]$.

7. $g_0 = g_0'$, $g_1 \neq g_1'$, $t_0 = t_0'$, and $t_1 \neq t_1'$: $C = N \cdot \mathbb{E}[\hat{\mu}^p_{g_0,g_1,t_0,t_1} \cdot \hat{\mu}^p_{g_0,g_1',t_0,t_1'}]$.

8. $g_0 = g_0'$, $g_1 \neq g_1'$, $t_0 \neq t_0'$, and $t_1 = t_1'$: $C = N \cdot \mathbb{E}[\hat{\mu}^q_{g_0,g_1,t_0,t_1} \cdot \hat{\mu}^q_{g_0,g_1',t_0',t_1}]$.

9. $g_0 = g_0'$, $g_1 \neq g_1'$, $t_0 \neq t_0'$, $t_1 \neq t_1'$, and $t_0' = t_1$: $C = N \cdot \mathbb{E}[\hat{\mu}^q_{g_0,g_1,t_0,t_1} \cdot \hat{\mu}^p_{g_0,g_1',t_1,t_1'}]$.

10. $g_0 = g_0'$, $g_1 \neq g_1'$, $t_0 \neq t_0'$, $t_1 \neq t_1'$, and $t_0 = t_1'$: $C = N \cdot \mathbb{E}[\hat{\mu}^p_{g_0,g_1,t_0,t_1} \cdot \hat{\mu}^q_{g_0,g_1',t_0',t_0}]$.

11. $g_0 \neq g_0'$, $g_1 = g_1'$, $t_0 = t_0'$, and $t_1 = t_1'$: $C = N \cdot \mathbb{E}[\hat{\mu}^r_{g_0,g_1,t_0,t_1} \cdot \hat{\mu}^r_{g_0',g_1,t_0,t_1}] + N \cdot \mathbb{E}[\hat{\mu}^s_{g_0,g_1,t_0,t_1} \cdot \hat{\mu}^s_{g_0',g_1,t_0,t_1}]$.

12. $g_0 \neq g_0'$, $g_1 = g_1'$, $t_0 = t_0'$, and $t_1 \neq t_1'$: $C = N \cdot \mathbb{E}[\hat{\mu}^r_{g_0,g_1,t_0,t_1} \cdot \hat{\mu}^r_{g_0',g_1,t_0,t_1'}]$.

13. $g_0 \neq g_0'$, $g_1 = g_1'$, $t_0 \neq t_0'$, and $t_1 = t_1'$: $C = N \cdot \mathbb{E}[\hat{\mu}^s_{g_0,g_1,t_0,t_1} \cdot \hat{\mu}^s_{g_0',g_1,t_0',t_1}]$.

14. $g_0 \neq g_0'$, $g_1 = g_1'$, $t_0 \neq t_0'$, $t_1 \neq t_1'$, and $t_0' = t_1$: $C = N \cdot \mathbb{E}[\hat{\mu}^s_{g_0,g_1,t_0,t_1} \cdot \hat{\mu}^r_{g_0',g_1,t_1,t_1'}]$.

15. $g_0 \neq g_0'$, $g_1 = g_1'$, $t_0 \neq t_0'$, $t_1 \neq t_1'$, and $t_0 = t_1'$: $C = N \cdot \mathbb{E}[\hat{\mu}^r_{g_0,g_1,t_0,t_1} \cdot \hat{\mu}^s_{g_0',g_1,t_0',t_0}]$.

16. $g_0 \neq g_0'$, $g_1 \neq g_1'$, $g_0' = g_1$, $t_0 = t_0'$, and $t_1 = t_1'$: $C = N \cdot \mathbb{E}[\hat{\mu}^r_{g_0,g_1,t_0,t_1} \cdot \hat{\mu}^p_{g_1,g_1',t_0,t_1}] + N \cdot \mathbb{E}[\hat{\mu}^s_{g_0,g_1,t_0,t_1} \cdot \hat{\mu}^q_{g_1,g_1',t_0,t_1}]$.

17. $g_0 \neq g_0'$, $g_1 \neq g_1'$, $g_0' = g_1$, $t_0 = t_0'$, and $t_1 \neq t_1'$: $C = N \cdot \mathbb{E}[\hat{\mu}^r_{g_0,g_1,t_0,t_1} \cdot \hat{\mu}^p_{g_1,g_1',t_0,t_1'}]$.

18. $g_0 \neq g_0'$, $g_1 \neq g_1'$, $g_0' = g_1$, $t_0 \neq t_0'$, and $t_1 = t_1'$: $C = N \cdot \mathbb{E}[\hat{\mu}^s_{g_0,g_1,t_0,t_1} \cdot \hat{\mu}^q_{g_1,g_1,t_0',t_1}]$.

19. $g_0 \neq g_0'$, $g_1 \neq g_1'$, $g_0' = g_1$, $t_0 \neq t_0'$, $t_1 \neq t_1'$, and $t_0' = t_1$: $C = N \cdot \mathbb{E}[\hat{\mu}^s_{g_0,g_1,t_0,t_1} \cdot \hat{\mu}^p_{g_1,g_1',t_1,t_1'}]$.

20. $g_0 \neq g_0'$, $g_1 \neq g_1'$, $g_0' = g_1$, $t_0 \neq t_0'$, $t_1 \neq t_1'$, and $t_0 = t_1'$: $C = N \cdot \mathbb{E}[\hat{\mu}^r_{g_0,g_1,t_0,t_1} \cdot \hat{\mu}^q_{g_1,g_1',t_0',t_0}]$.

21. $g_0 \neq g_0'$, $g_1 \neq g_1'$, $g_0 = g_1'$, $t_0 = t_0'$, and $t_1 = t_1'$: $C = N \cdot \mathbb{E}[\hat{\mu}^p_{g_0,g_1,t_0,t_1} \cdot \hat{\mu}^s_{g_0',g_0,t_0,t_1}] + N \cdot \mathbb{E}[\hat{\mu}^q_{g_0,g_1,t_0,t_1} \cdot \hat{\mu}^r_{g_0',g_0,t_0,t_1}]$.

22. $g_0 \neq g_0'$, $g_1 \neq g_1'$, $g_0 = g_1'$, $t_0 = t_0'$, and $t_1 \neq t_1'$: $C = N \cdot \mathbb{E}[\hat{\mu}^p_{g_0,g_1,t_0,t_1} \cdot \hat{\mu}^s_{g_0',g_0,t_0,t_1'}]$.

23. $g_0 \neq g_0'$, $g_1 \neq g_1'$, $g_0 = g_1'$, $t_0 \neq t_0'$, and $t_1 = t_1'$: $C = N \cdot \mathbb{E}[\hat{\mu}^q_{g_0,g_1,t_0,t_1} \cdot \hat{\mu}^r_{g_0',g_0,t_0',t_1}]$.

24. $g_0 \neq g_0'$, $g_1 \neq g_1'$, $g_0 = g_1'$, $t_0 \neq t_0'$, $t_1 \neq t_1'$, and $t_0' = t_1$: $C = N \cdot \mathbb{E}[\hat{\mu}^q_{g_0,g_1,t_0,t_1} \cdot \hat{\mu}^r_{g_0',g_0,t_1,t_1'}]$.

25. $g_0 \neq g_0'$, $g_1 \neq g_1'$, $g_0 = g_1'$, $t_0 \neq t_0'$, $t_1 \neq t_1'$, and $t_0 = t_1'$: $C = N \cdot \mathbb{E}[\hat{\mu}^p_{g_0,g_1,t_0,t_1} \cdot \hat{\mu}^s_{g_0',g_0,t_0',t_0}]$.

26. $g_0 \neq g_0'$, $g_1 \neq g_1'$, $g_0 \neq g_1'$, and $g_0' \neq g_1$: $C = 0$.

REFERENCES

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

——— (2005): "Semiparametric Difference-in-Differences Estimators," *Review of Economic Studies*, 72, 1–19.

ABADIE, A., J. ANGRIST, AND G. IMBENS (2002): "Instrumental Variables Estimates of the Effect of Training on the Quantiles of Trainee Earnings," *Econometrica*, 70, 91–117.

ALTONJI, J., AND R. BLANK (2000): "Race and Gender in the Labor Market," in *Handbook of Labor Economics*, ed. by O. Ashenfelter and D. Card. Amsterdam: Elsevier, 3143–3259.

ALTONJI, J., AND R. MATZKIN (1997): "Panel Data Estimators for Nonseparable Models with Endogenous Regressors," Mimeo, Department of Economics, Northwestern University.

——— (2005): "Cross-Section and Panel Data Estimators for Nonseparable Models with Endogenous Regressors," *Econometrica*, 73, 1053–1102.

ANGRIST, J., AND A. KRUEGER (2000): "Empirical Strategies in Labor Economics," in *Handbook of Labor Economics*, ed. by O. Ashenfelter and D. Card. Amsterdam: Elsevier, 1277–1366.

ASHENFELTER, O., AND D. CARD (1985): "Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs," *Review of Economics and Statistics*, 67, 648–660.

ASHENFELTER, O., AND M. GREENSTONE (2004): "Using the Mandated Speed Limits to Measure the Value of a Statistical Life," *Journal of Political Economy*, 112, S226–S267.

ATHEY, S., AND G. IMBENS (2002): "Identification and Inference in Nonlinear Difference-in-Differences Models," Technical Working Paper t0280, National Bureau of Economic Research.

——— (2006): "Supplement to 'Identification and Inference in Nonlinear Difference-in-Difference Models'," *Econometrica Supplementary Material*, Vol. 74, http://econometricsociety.org/ecta/supmat/4035extensions.pdf.

ATHEY, S., AND S. STERN (2002): "The Impact of Information Technology on Emergency Health Care Outcomes," *RAND Journal of Economics*, 33, 399–432.

BARNOW, B. S., G. G. CAIN, AND A. S. GOLDBERGER (1980): "Issues in the Analysis of Selectivity Bias," in *Evaluation Studies*, Vol. 5, ed. by E. Stromsdorfer and G. Farkas. San Francisco: Sage, 43–59.

BERTRAND, M., E. DUFLO, AND S. MULLAINATHAN (2004): "How Much Should We Trust Differences-in-Differences Estimates?" *Quarterly Journal of Economics*, 119, 249–275.

BESLEY, T., AND A. CASE (2000): "Unnatural Experiments? Estimating the Incidence of Endogenous Policies," *Economic Journal*, 110, F672–694.

BILLINGSLEY, P. (1968): *Probability and Measure* (Second Ed.). New York, Wiley.

BLUNDELL, R., M. COSTA DIAS, C. MEGHIR, AND J. VAN REENEN (2001): "Evaluating the Employment Impact of a Mandatory Job Search Assistance Program," Working Paper 01/20, IFS, University College London.

BLUNDELL, R., A. DUNCAN, AND C. MEGHIR (1998): "Estimating Labour Supply Responses Using Tax Policy Reforms," *Econometrica*, 66, 827–861.

BLUNDELL, R., AND T. MACURDY (2000): "Labor Supply," in *Handbook of Labor Economics*, ed. by O. Ashenfelter and D. Card. Amsterdam: Elsevier, 1559–1695.

BORENSTEIN, S. (1991): "The Dominant-Firm Advantage in Multiproduct Industries: Evidence from the U.S. Airlines," *Quarterly Journal of Economics*, 106, 1237–1266.

CARD, D. (1990): "The Impact of the Mariel Boatlift on the Miami Labor Market," *Industrial and Labor Relations Review*, 43, 245–257.

CARD, D., AND A. KRUEGER (1993): "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania," *American Economic Review*, 84, 772–784.

CHAY, K., AND D. LEE (2000): "Changes in the Relative Wages in the 1980s: Returns to Observed and Unobserved Skills and Black–White Wage Differentials," *Journal of Econometrics*, 99, 1–38.

CHERNOZHUKOV, V., AND C. HANSEN (2005): "An IV Model of Quantile Treatment Effects," *Econometrica*, 73, 245–261.

CHESHER, A. (2003): "Identification in Nonseparable Models," *Econometrica*, 71, 1405–1441.

CHIN, A. (2005): "Long-Run Labor Market Effects of the Japanese-American Internment During World War II," *Journal of Labor Economics*, 23, 491–525.

DAS, M. (2001): "Monotone Comparative Statics and the Estimation of Behavioral Parameters," Working Paper, Department of Economics, Columbia University.

——— (2004): "Instrumental Variables Estimators for Nonparametric Models with Discrete Endogenous Regressors," *Journal of Econometrics*, 124, 335–361.

DEHEJIA, R. (1997): "A Decision-Theoretic Approach to Program Evaluation," Ph.D. Dissertation, Department of Economics, Harvard University.

DEHEJIA, R., AND S. WAHBA (1999): "Causal Effects in Non-Experimental Studies: Re-Evaluating the Evaluation of Training Programs," *Journal of the American Statistical Association*, 94, 1053–1062.

DONALD, S., AND K. LANG (2001): "Inference with Difference in Differences and Other Panel Data," Unpublished Manuscript, Boston University.

DONOHUE, J., J. HECKMAN, AND P. TODD (2002): "The Schooling of Southern Blacks: The Roles of Legal Activism and Private Philanthropy, 1910–1960," *Quarterly Journal of Economics*, 117, 225–268.

DUFLO, E. (2001): "Schooling and Labor Market Consequences of School Construction in Indonesia: Evidence from an Unusual Policy Experiment," *American Economic Review*, 91, 795–813.

EISSA, N., AND J. LIEBMAN (1996): "Labor Supply Response to the Earned Income Tax Credit," *Quarterly Journal of Economics*, 111, 605–637.

FORTIN, N., AND T. LEMIEUX (1999): "Rank Regressions, Wage Distributions and the Gender Gap," *Journal of Human Resources*, 33, 611–643.

GRUBER, J., AND B. MADRIAN (1994): "Limited Insurance Portability and Job Mobility: The Effects of Public Policy on Job-Lock," *Industrial and Labor Relations Review*, 48, 86–102.

HAHN, J. (1998): "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica*, 66, 315–331.

HECKMAN, J. (1996): "Discussion," in *Empirical Foundations of Household Taxation*, ed. by M. Feldstein and J. Poterba. Chicago: University of Chicago Press, 32–38.

HECKMAN, J. J., AND B. S. PAYNER (1989): "Determining the Impact of Federal Antidiscrimination Policy on the Economic Status of Blacks: A Study of South Carolina," *American Economic Review*, 79, 138–177.

HECKMAN, J., AND R. ROBB (1985): "Alternative Methods for Evaluating the Impact of Interventions," in *Longitudinal Analysis of Labor Market Data*, ed. by J. Heckman and B. Singer. New York: Cambridge University Press, 156–245.

HIRANO, K., G. IMBENS, AND G. RIDDER (2003): "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica*, 71, 1161–1189.

HONORE, B. (1992): "Trimmed LAD and Least Squares Estimation of Truncated and Censored Regression Models with Fixed Effects," *Econometrica*, 63, 533–565.

IMBENS, G., AND J. ANGRIST (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467–475.

IMBENS, G. W., AND C. F. MANSKI (2004): "Confidence Intervals for Partially Identified Parameters," *Econometrica*, 72, 1845–1857.

IMBENS, G., AND W. NEWEY (2001): "Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity," Mimeo, Department of Economics, UC Berkeley and MIT.

JIN, G., AND P. LESLIE (2003): "The Effect of Information on Product Quality: Evidence from Restaurant Hygiene Grade Cards," *Quarterly Journal of Economics*, 118, 409–451.

JUHN, C., K. MURPHY, AND B. PIERCE (1991): "Accounting for the Slowdown in Black–White Wage Convergence," in *Workers and Their Wages*, ed. by M. Kosters. Washington, DC: AEI Press, 107–143.

——— (1993): "Wage Inequality and the Rise in Returns to Skill," *Journal of Political Economy*, 101, 410–442.

KRUEGER, A. (1999): "Experimental Estimates of Education Production Functions," *Quarterly Journal of Economics*, 114, 497–532.

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

LECHNER, M. (1999): "Earnings and Employment Effects of Continuous Off-the-Job Training in East Germany after Unification," *Journal of Business & Economic Statistics*, 17, 74–90.

MANSKI, C. (1990): "Non-Parametric Bounds on Treatment Effects," *American Economic Review, Papers and Proceedings*, 80, 319–323.

——— (1995): *Identification Problems in the Social Sciences*. Cambridge, MA: Harvard University Press.

MARRUFO, G. (2001): "The Incidence of Social Security Regulation: Evidence from the Reform in Mexico," Mimeo, University of Chicago.

MATZKIN, R. (1999): "Nonparametric Estimation of Nonadditive Random Functions," Mimeo, Department of Economics, Northwestern University.

——— (2003): "Nonparametric Estimation of Nonadditive Random Functions," *Econometrica*, 71, 1339–1375.

MEYER, B. (1995): "Natural and Quasi-Experiments in Economics," *Journal of Business & Economic Statistics*, 13, 151–161.

MEYER, B., K. VISCUSI, AND D. DURBIN (1995): "Workers' Compensation and Injury Duration: Evidence from a Natural Experiment," *American Economic Review*, 85, 322–340.

MOFFITT, R., AND M. WILHELM (2000): "Taxation and the Labor Supply Decisions of the Affluent," in *Does Atlas Shrug? Economic Consequences of Taxing the Rich*, ed. by Joel Slemrod. Cambridge, MA: Harvard University Press, 193–234.

MOULTON, B. R. (1990): "An Illustration of a Pitfall in Estimating the Effects of Aggregate Variables on Micro Unit," *Review of Economics and Statistics*, 72, 334–338.

NEWEY, W. (1994): "The Asymptotic Variance of Semiparametric Estimators," *Econometrica*, 62, 1349–1382.

POTERBA, J., S. VENTI, AND D. WISE (1995): "Do 401(k) Contributions Crowd Out Other Personal Saving?" *Journal of Public Economics*, 58, 1–32.

ROSENBAUM, P., AND D. RUBIN (1983): "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41–55.

RUBIN, D. (1974): "Estimating Causal Effects of Treatments in Randomized and Non-Randomized Studies," *Journal of Educational Psychology*, 66, 688–701.

——— (1978): "Bayesian Inference for Causal Effects: The Role of Randomization," *The Annals of Statistics*, 6, 34–58.

SHADISH, W., T. COOK, AND D. CAMPBELL (2002): *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin.

SHORACK, G., AND J. WELLNER (1986): *Empirical Processes with Applications to Statistics*. New York: Wiley.

STUTE, W. (1982): "The Oscillation Behavior of Empirical Processes," *The Annals of Probability*, 10, 86–107.

VAN DER VAART, A. (1998), *Asymptotic Statistics*. Cambridge, U.K.: Cambridge University Press.