Prognostic Score-Based Difference-in-Differences Strategy

Guanglei Hong[1], Takako Nomi[2], Bing Yu[1]
[1]University of Chicago, 5736 S Woodlawn Ave, Chicago, IL 60637
[2]Saint Louis University, 3500 Lindell Blvd, Fitzgerald Hall, Saint Louis, MO 63103

Abstract
In policy evaluations, the standard difference-in-differences (DID) method relies on the strong assumption that the average confounding effect of concurrent events is the same for the comparison group unaffected by the policy and the experimental group affected by the policy. Recent advancements include using propensity score matching or weighting to equate the covariate distribution between the comparison group and the experimental group. Another approach is to estimate the distribution of the counterfactual outcome of the experimental group resembling the outcome change in the comparison group. We propose an alternative strategy that involves a pair of prognostic scores per unit representing the predicted pre-policy outcome and the predicted post-policy outcome under the comparison condition in the absence of policy change. Subsequent DID analyses within subclasses defined by this pair of prognostic scores allow for a calibrated adjustment. This study compares the identification assumptions required by the prognostic score-based strategy with those of the existing strategies. We illustrate with an evaluation of a policy requiring all ninth graders to take algebra.

Key Words: Causal inference, econometrics, longitudinal data, natural experiment, policy analysis, propensity score, quasi-experimental design

When using time series data to evaluate system-wide policies, concurrent changes often pose threats to internal validity. The standard difference-in-differences (DID) method resorts to a non-equivalent comparison group whose average outcome change is due to such confounding. This strategy relies on the strong assumption that the average confounding impact of concurrent events is the same for the comparison group unaffected by the policy and the experimental group affected by the policy. This assumption will be violated and therefore the DID results will be biased, for example, if the confounding effect varies by individual characteristics and if the experimental group and the comparison group differ in such characteristics.

In the recent econometrics literature, researchers have attempted to equate the covariate distribution of the comparison group with that of the experimental group in each time period through propensity score matching or weighting before conducting DID analyses (Abadie, 2005; Cerdá, et al, 2012; Heckman, Ichimura, Smith, & Todd, 1998; Heckman, Ichimura, & Todd, 1997). Another approach is to nonlinearly estimate the entire distribution of the post-policy outcome of the experimental group associated with the counterfactual absence of the policy resembling the change in the outcome distribution of the comparison group (Athey & Imbens, 2006). Each of these strategies invokes a set of strong assumptions that may not hold in a particular application.

We propose an alternative strategy that extends the Peters-Belson method (Belson, 1956; Peters, 1941) to the DID context. The use of prognostic scores (Hansen, 2008) in the causal inference literature can be viewed as the latest development of the Peters-Belson method. Specifically, our strategy involves a pair of prognostic scores per unit representing the predicted pre-policy outcome and the predicted post-policy outcome

if unaffected by the policy. By conducting DID analyses within subclasses defined by this pair of prognostic scores, the adjustment is calibrated in relation to the covariates predicting the outcomes. Our rationale is to equate the predicted amount of confounding of concurrent events across the pre-policy experimental group, the post-policy experimental group, the pre-policy comparison group, and the post-policy comparison group within subclasses of units.

This study compares the identification assumptions required by the prognostic score-based strategy with the existing strategies. We clarify conditions under which the prognostic score-based DID solution is hypothesized to outperform the standard DID and the propensity score-based or nonlinear distribution-based strategies. We illustrate with an evaluation of a policy adopted by the Chicago Public Schools requiring all ninth graders to take algebra. The CPS data are representative of education accountability data collected by many school systems around the U.S. The theoretical results presented in this paper apply to a continuous outcome such as student achievement data as well as a binary outcome such as whether a student eventually graduates from high school.

The paper is organized as follows: section 1 describes the motivating example of a system-wide policy; section 2 introduces our notation and defines the causal estimands; section 3 derives the bias associated with the confounding of concurrent events in an unadjusted analysis, lists the key identification assumptions required by the standard DID analysis, and describes some major challenges to these assumptions in the current application; section 4 reviews the existing alternative DID strategies, clarifies their identification assumptions, and reveals their limitations when applied to the current example; section 5 introduces the prognostic score-based DID strategy and provides its theoretical rationale, identification assumptions, and analytic procedure; section 6 discusses the relative strengths and limitations of the prognostic score-based DID solution in comparison with the existing DID methods. We also discuss extensions of the prognostic score-based DID strategy to multilevel multi-cohort data.

1. Motivating Example

The algebra-for-all policy was adopted by the Chicago Public Schools in 1997. Prior to that year, whether a 9^{th} grader took algebra mostly depended on the student's math preparation in the elementary school. The algebra-for-all policy was intended to eliminate remedial math courses for low-achieving students and thereby improving high school math achievement across the board. Earlier research has shown a considerable amount of improvement in 9^{th} graders' average math achievement in the post-policy years in comparison with that in the pre-policy years.

However, CPS students experienced a number of important policy changes during those same years. The concurrent events include a policy retaining low-achieving students in 3^{rd}, 6^{th}, and 8^{th} grades, a change in the special education program, and an overall improvement in elementary education, all of which might lead to a change in 9^{th} graders' incoming math skills. Additionally, these other policies might change student response to the math curriculum. For example, once low-achieving students were retained and once special education students were provided with extra support, the 9^{th} graders in regular education might achieve a higher level of math learning due to the change in peer composition even without math curricular change. Concurrent policies might also change school personnel or school climate, which would then affect the quality of instruction and subsequently student learning. Hence the impact of replacing remedial math with algebra was likely confounded by the impacts of these concurrent interventions.

Among the 59 neighborhood high schools in Chicago that existed both before and after 1997, 45 schools offered remedial math to low-achieving students prior to 1997 and replaced remedial math with algebra after 1997; 14 schools offered algebra to all 9^{th}

graders prior to 1997 and thus were unaffected by the policy. This provides a possibility of using the DID strategy to remove the confounding of concurrent events. However, the standard DID strategy would produce biased results, for example, if the confounding effect is different for whites and minority students and if the racial composition differs between the two types of schools.

The current application involves repeated cross-sectional data of individuals (9th graders) nested in a panel of clusters (high schools). The constitution of student cohorts changed over time. Also available are repeated assessments of students prior to the 9th grade. Unlike most DID studies in which individuals in the experimental group started to experience the policy at a certain time while those in the comparison group never experienced the policy, in this application, the policy had already been implemented in the comparison schools before it was adopted by the experimental schools.

2. Notation and Causal Estimands

For simplicity, we start by focusing on the mean difference in the 9th grade math outcome between a pre-policy cohort and a post-policy cohort, contrasting a hypothetical experimental school with a hypothetical comparison school. We will then extend the results to multiple schools and multi-cohort time series data.

Let Y_i denote the math outcome of student i at the end of the 9th grade measured on a continuous scale. Let $G_i = 1$ if the student attended an experimental school affected by the policy; let $G_i = 0$ if the student attended a comparison school unaffected by the policy. Let $T_i = 1$ if the student was enrolled in the 9th grade during the post-policy year and 0 if the student was enrolled in the pre-policy year. Let X_i denote a vector of covariates measuring student characteristics that are not affected by the policy.

Let $Y_{iG1.T1}^{(1)}$ denote the potential outcome that student *i* would display if attending the experimental school and having exposure to the policy in the post-policy year; let $Y_{iG1.T1}^{(0)}$ denote the student's potential outcome if the student in the experimental school counterfactually had no exposure to the policy in the post-policy year should the policy have been postponed. Here the superscript indicates policy exposure while the subscript indicates school membership and cohort membership. Suppose that we are interested in estimating the average policy effect for students attending the experimental school in the post-policy year (i.e., the treatment effect on the treated in the experimental group). The causal estimand is

$$\delta_{G1.T1} = E\left[Y_{G1.T1}^{(1)} - Y_{G1.T1}^{(0)} | G = 1, T = 1\right].$$

If, instead, we are interested in estimating the average policy effect for students attending the experimental school in the pre-policy year (i.e., the treatment effect on the untreated in the experimental group), we may consider the possibility that the policy could have been introduced in an earlier year. Let $Y_{iG1.T0}^{(1)}$ denote the potential outcome that student i would display if attending the experimental school and counterfactually having exposure to the policy in the pre-policy year; let $Y_{iG1.T0}^{(0)}$ denote the student's potential outcome in the pre-policy year in the absence of the policy. The causal estimand becomes

$$\delta_{G1.T0} = E\left[Y_{G1.T0}^{(1)} - Y_{G1.T0}^{(0)} | G = 1, T = 0\right].$$

If the pre-policy cohort and the post-policy cohort in the experimental school differ in covariate distribution that is not caused by the policy and if the policy effect depends on covariate values, then the average policy effect may differ between the two cohorts. In the current application, we focus on the policy impact on the pre-policy cohort $\delta_{G1.T0}$ in which $Y_{G1.T0}^{(1)}$ is counterfactual for those attending the experimental school in the

pre-policy year. Because the comparison group in this study had exposure to the policy in both years, it allows us to estimate the amount of confounding of concurrent events as the difference between $Y_{G1.T1}^{(1)}$ and $Y_{G1.T0}^{(1)}$, which then makes possible the estimation of $\delta_{G1.T0}$.

3. Bias due to Concurrent Events and the Standard DID Method

For pre-policy students in the experimental school, in order to estimate $\delta_{G1.T0}$, a naive analysis would use the observed outcome information from the post-policy cohort $Y_{G1.T1}^{(1)}$ to identify the counterfactual outcome of the pre-policy cohort $Y_{G1.T0}^{(1)}$. The estimation would be unbiased under the assumption that

$$E\left[Y_{G1.T1}^{(1)}|G=1,T=1\right]=E\left[Y_{G1.T0}^{(1)}|G=1,T=0\right].$$

That is, the pre-policy cohort and the post-policy cohort in the experimental school would have displayed the same mean outcome if both cohorts had been exposed to the policy. However, even if these two cohorts in the experimental school were identical, the above assumption would still be untenable due to overwhelming evidence for the confounding of concurrent events that we have mentioned in section 1.

The average effect of the concurrent events for students in the experimental school can be defined as follows:

$$b_{G1}=E\left[Y_{G1.T1}^{(1)}|G=1,T=1\right]-E\left[Y_{G1.T0}^{(1)}|G=1,T=0\right].$$

The comparison school had the policy in both years. Thus, the average confounding effect for the comparison school students is

$$b_{G0}=E\left[Y_{G0.T1}^{(1)}|G=0,T=1\right]-E\left[Y_{G0.T0}^{(1)}|G=0,T=0\right].$$

The standard DID method removes the confounding effect under the assumption

$$b_{G1}=b_{G0}. \quad (1)$$

Let D_{G1} denote the observed difference between the pre-policy cohort and the post-policy cohort in the experimental school:

$$D_{G1}=E[Y|G=1,T=1]-E[Y|G=1,T=0].$$

Let D_{G0} denote the observed difference between the pre-policy cohort and the post-policy cohort in the comparison school:

$$D_{G0}=E[Y|G=0,T=1]-E[Y|G=0,T=0].$$

The standard DID estimator is $D_{G1}-D_{G0}$, which estimates the causal estimand $\delta_{G1.T0}$ when assumption (1) holds. One obtains the standard DID estimator through analyzing a linear model

$$Y_i=\alpha_0+\alpha_1T_i+\beta_0G_i+\beta_1G_iT_i+e_i. \quad (2)$$

However, when the key assumption in equation (1) does not hold, the standard DID method will lead to bias in an amount equal to $b_{G1}-b_{G0}$. Assumption (1) will always hold when the confounding effect of concurrent events is a constant for all units in the population. However, when there is heterogeneity in the confounding effect, the standard DID estimate is biased under the following two scenarios. Here we let *X* denote observed pretreatment covariates and *U* for unobserved pretreatment covariates. They represent individual characteristics that cannot be affected by the policy.

Scenario 1: Assumption (1) is violated if the confounding effects of concurrent events are heterogeneous with respect to *X* or *U* and if the experimental school and the comparison school differ in the composition of *X* or *U* during the pre-policy year (Meyer, 1995). For example, if racial composition differs between the experimental school and the comparison school in the pre-policy year and if the confounding factors have differential effects on whites and minority students, the standard DID estimate will be biased.

Scenario 2: Even if the experimental school and the comparison school had the same distribution of X and U in the pre-policy year, they may experience different historical changes in X or U in the absence of the policy change, which would lead to a difference in pretreatment composition between these two schools during the post-policy year. If the confounding effects of concurrent events are heterogeneous with respect to X or U, the standard DID will again produce a biased result.

4. Existing Alternative DID Strategies

4.1. DID with Linear Covariance Adjustment

Prior research has typically employed the following DID model with linear covariance adjustment for observed pretreatment characteristics X (e.g., Barnow, Cain, & Goldberger, 1980; Card & Kruger, 1993; Dynarski, 2003; Fitzpatrick, 2008). Under model-based assumptions, this strategy may adjust for (a) differences between the experimental group and the comparison group in X in the pre-policy year as well as (b) additional between-group differences in X in the post-policy year that is not a result of the policy. In addition to bias removal, inclusion of X may improve the precision of the policy effect estimate by reducing error variance.

$$Y_i = \alpha_0 + \alpha_1 T_i + \beta_0 G_i + \beta_1 G_i T_i + \lambda X_i + e_i \quad (3)$$

Here β_1 is an unbiased estimate of the causal estimand $\delta_{G1.T0}$ under a key assumption, that is, the confounding effect of concurrent events for the experimental group is the same as that for the comparison group within levels of pretreatment covariates $X = x$:

$$b_{G1|X} = b_{G0|X} \quad (4)$$

where

$$b_{G1|X} = E\left(Y^{(1)}_{G1.T1} | G = 1, T = 1, X = x\right) - E\left(Y^{(1)}_{G1.T0} | G = 1, T = 0, X = x\right),$$
$$b_{G0|X} = E\left(Y^{(1)}_{G0.T1} | G = 0, T = 1, X = x\right) - E\left(Y^{(1)}_{G0.T0} | G = 0, T = 0, X = x\right).$$

The following assumptions are additionally implied by Model (3):

(5a) Conditioning on X, the average treatment effect for the untreated in the experimental group is the same as that for the entire population should all units have been untreated. This assumption is required because the above integral is taken over the marginal distribution of X rather than the conditional distribution of X given $G = 1$ and $T = 0$.

(5b) The functional form of the outcome model is correctly specified. This assumption is critical when the experimental group and the comparison group do not have complete overlap in the distribution of X such that inference is based on extrapolation.

(5c) The X-Y relationships are the same between the experimental group and the comparison group and are invariant across time.

4.2. DID with Propensity Score Adjustment

Heckman and colleagues (1997, 1998) used propensity score matching to identify the common support in the observed covariates X and to equate the distribution of X between the experimental group and the comparison group. Abadie (2005) proposed using propensity score-based inverse-probability-of-treatment weighting to equate the distribution of X between the two groups. Analyzing longitudinal data in which a single cohort of students experienced no change in policy over time if in the comparison group and experienced a change in policy if in the experimental group, Heckman and colleagues implicitly assumed that the pretreatment composition of each group does not change over time. For modeling repeated cross-sectional data, Abadie (2005) alternatively assumed that, within a treatment group, the pre-policy observations and the post-policy observations are random samples from the same population, and therefore the

pretreatment composition is expected to be the same between the pre-policy cohort and the post-policy cohort.

The propensity score-based methods do not rely on outcome model specifications. They thereby avoid assumptions (5a), (5b), and (5c) required by linear covariance adjustment. Assumption (4) remains necessary. That is, the confounding effect of concurrent events for the experimental group should be the same as that for the comparison group on average within levels of observed covariates *X* (see Heckman et al, 1997, 1998, and Abadie 2005). In other words, given the observed pretreatment covariates *X*, there should be no unobserved differences in the confounding between the experimental group and the comparison group. Yet in carrying out the propensity score-based methods, the researchers invoked an assumption stronger than assumption (4). Let $\phi(X) = pr(G = 1|X)$ be the propensity score representing the conditional probability that an individual would be assigned to the experimental group given *X*. Propensity score matching or inverse-probability-of-treatment weighting assumes the following:

$$Y_{G1.T1}^{(1)}, Y_{G0.T1}^{(1)} \perp G \left| T = 1, \phi(X); \text{ and } Y_{G1.T0}^{(1)}, Y_{G0.T0}^{(1)} \perp G \right| T = 0, \phi(X). \quad (6)$$

Assumption (6) is much stronger than assumption (4). This is because assumption (6) requires $E\left(Y_{G1.T1}^{(1)}|G = 1, T = 1, X = x\right) = E\left(Y_{G0.T1}^{(1)}|G = 0, T = 1, X = x\right)$ and $E\left(Y_{G1.T0}^{(1)}|G = 1, T = 0, X = x\right) = E\left(Y_{G0.T0}^{(1)}|G = 0, T = 0, X = x\right)$. In contrast, assumption (4) holds even when the average potential outcomes of the experimental group and the comparison group are unequal in a given year.

Applying the propensity score-based matching method, one would analyze a propensity score model for being assigned to the experimental school, and match the comparison school students to those in the experimental school on the basis of the estimated propensity score. One would then apply the standard DID model to the matched data. The rationale and assumptions of propensity score weighting adjustment are similar to those of propensity score matching. To implement, one may compute an inverse-probability weight for each student as a function of the estimated propensity score. The standard DID model is then applied to the weighted data (Abadie, 2005).

In a study of repeated cross-sectional data in which the pretreatment composition of the two cohorts within each treatment group are presumably different, Blundell et al (2004) alternatively suggested using two pairs of propensity scores (one for the experimental versus the comparison group in a given time period and the other for time period given group membership) such that all four cells are balanced on observed characteristics. This alternative procedure assumes the same mechanism for treatment group selection across the pre-policy cohort and the post-policy cohort.

4.3. Nonlinear Changes-in-Changes (CIC) Adjustment

When the experimental group and the comparison group are different in unobserved pretreatment characteristics *U*, to estimate the treatment effect on the treated, Athey and Imbens (2006) proposed a nonlinear CIC model estimating the entire distribution of the counterfactual outcome for the experimental group based on the observed change in the outcome distribution of the comparison group. The assumptions were stated in the case that the comparison group was never exposed to the policy.

(7a) *A single index model and common change in production function between groups*. Holding policy constant, the outcome satisfies the production function $Y^{(0)} = h(u, t)$ for $t = 0, 1$ that applies to both the experimental group and the comparison group in a given time period. All outcome differences between the two groups are attributed to between-group differences in *U*. In the absence of policy change, the change over time in the outcome distribution of each group arises from the fact that $h(u, 0)$ in the pre-policy

year differs from $h(u,1)$ in the post-policy year due to concurrent confounding factors; the change from $h(u,0)$ to $h(u,1)$ is assumed to be common to the two groups.

(7b) *Strict monotonicity*. The production function $h(u,t)$ is strictly increasing in u given *t*.

(7c) *Time invariance within groups*: $U \perp T|G$. The population of units within a given group does not change over time and therefore the distribution of *U* does not change. In other words, *U* cannot be affected by concurrent confounding factors. Any differences between the experimental group and the comparison group in *U* therefore should be stable over time.

(7d) The support for the distribution of *U* in the experimental group is a subset of that in the comparison group: $\mathbb{U}_1 \subseteq \mathbb{U}_0$. When this assumption does not hold, the population of relevant interest will need to be re-defined.

Applying this strategy to our example where the comparison group was always exposed to the policy, we define the conditional outcome distribution in group *g* and time *t* as $F_{Y.gt}$. Intuitively, there is a u value corresponding to each outcome value in both groups in a given year. For an experimental unit whose observed post-policy outcome value is *y*, its counterpart in the comparison group would have the same post-policy outcome y determined by u through the monotonic function $h(u,1)$ that links the distribution of *U* with $F_{Y.01}$. Because the distribution of *U* is time invariant within each group, one then links u to an observed pre-policy outcome value y' in the comparison group through the function $h(u,0)$. Here y' is the counterfactual outcome associated with policy exposure for experimental units in the pre-policy year with characteristics *u*. In general, the counterfactual outcome $Y_{G1.T0}^{(1)}$ for an experimental unit with an unobserved component u such that $h(u,0) = y$ can be estimated by the CIC model: $k^{CIC}(y) = F_{Y.01}^{-1}(F_{Y.00}(y))$, where $F_Y^{-1}(q) = \inf\{y \in \mathbb{Y}: F_Y(y) \geq q\}$ for $q \in [0,1]$ is the inverse distribution function.

4.4. Strengths and Limitations of the Existing Alternative DID Strategies

Unlike the standard DID models represented by model (2), the existing alternative DID strategies apply when the average amount of confounding due to concurrent events differs between the experimental group and the comparison group. Below we summarize their relative strengths and limitations.

Linear Covariance Adjusted DID. DID models with linear covariance adjustment for observed pretreatment covariates *X* are easy to implement through ordinary least squares. This method is nonetheless constrained by the model specification. For example, model (3) would generate a biased estimate of the policy effect if, in addition to the differences in *X* between the experimental group and the comparison group during the pre-policy year or the post-policy year, the *X*-*Y* relationships are also different between the two groups within a cohort under the same policy. To relax assumption (5c) of invariant *X*-*Y* relationships, one may fit a DID model including higher-order interactions. Below is a saturated linear model including all possible interactions:

$$Y_i = \alpha_0 + \alpha_1 T_i + \beta_0 G_i + \beta_1 G_i T_i + \theta_{00} X_i + \alpha_2 T_i X_i + \beta_2 G_i X_i + \beta_3 G_i T_i X_i + e_i.$$

Here β_3 indicates whether the policy effect linearly depends on X. However, as the number of covariates increases and as nonlinearity arises, model misspecification becomes increasingly likely and is consequential for identification. Additionally, linear covariance adjusted DID models can lead to extrapolations outside the allowable range.

Propensity Score-Based DID. The propensity score-based DID methods seem suitable when there is a large dimension of observed pretreatment covariates. These methods allow the policy effect to be heterogeneous across different levels of propensity scores without explicitly modeling policy interactions with covariates. Yet the propensity

score-based strategies make no use of covariates that are unrelated to group membership *G* but are predictive of *Y*. For example, student incoming math skills may show comparable distributions between the experimental school and the comparison school such that the measure of incoming skills plays no role in the propensity score model for *G*. However, adjusting for incoming skills would potentially improve the precision of the policy estimate especially if the confounding of concurrent events is a function of incoming skills. In the meantime, propensity score models may include covariates that are related to *G* but are unrelated to *Y*. Adjusting for such covariates would likely reduce precision due to a reduction in common support and may even introduce bias when propensity score-based weighting is employed (Austin, 2008; Brookhart, Schneeweiss, Rothman, Glynn, Avorn, & Stürmer, 2006). More importantly, the propensity score-based methods invoke assumption (6) that is considerably stronger than necessary. DID analyses do not require that the comparison group and the experimental group have the same pretreatment composition. Rather, it only requires that the two groups or subsets of units from the two groups be affected in the same amount by concurrent confounding factors. Hence the range of common support for a DID analysis might be considerably larger than one might obtain when using propensity scores.

Nonlinear CIC Model. The nonlinear CIC models show particular strengths when the experimental group and the comparison group are different in the distribution of an unobservable. They require no assumptions with regard to the functional form of the models and are invariant to the scaling of the outcomes. This strategy also allows for heterogeneous policy effect without explicitly modeling the heterogeneity. This is because the amount of confounding $h(u,1) - h(u,0)$ may vary across units. The CIC models can also be extended to the case with observed covariates *X* if all the assumptions hold conditional on *X* (Athey and Imbens, 2006). However, the nonlinear CIC method requires strong assumptions that may not always be plausible. For example, when the distribution of *U* differs between the experimental group and the comparison group in a given time period under the same policy, it seems likely that the *U*-*Y* relationship may differ between the two groups as well, which will invalidate Assumption (7a). The method requires that *U* explains all the variation in the outcome, which rules out classical measurement error in the outcome. This seems unrealistic because measurement error is often inevitable. In standardized educational tests, the reliability is typically around .95, implying 5% of the error variance in the student outcome. It may also be rare for assumption (7b) strict monotonicity to hold for the relationship between *Y* and *U*. Besides, the nonlinear CIC method becomes inapplicable when the distribution of *U* differs across repeated cross-sectional cohorts.

In light of the limitations of the existing DID strategies, we propose a prognostic score-based DID alternative that relies on a different set of assumptions arguably more plausible in the current application.

5. Prognostic Score-Based DID Strategy

5.1. Theoretical Rationale

We have noted earlier that assumption (4) is sufficient for estimating $\delta_{G1.T0}$, the policy effect of interest. The rationale for using prognostic scores is that, rather than attempting to equate the covariate distribution between the experimental school and the comparison school in terms of how likely a student would select the experimental school as is the goal of the propensity score model, we attempt to directly estimate the amount of confounding associated with concurrent events if a student would attend the comparison school. Under Assumption (4), this is assumed equal to the amount of confounding if the

student would attend the experimental school. In other words, across all four *G*-by-*T* groups of students, we attempt to identify those who share the following quantity:

$$b_{G0|X} = E\left[Y_{G0.T1}^{(1)}|X\right] - E\left[Y_{G0.T0}^{(1)}|X\right].$$

Hence our goal is to identify a subpopulation of students defined by $X = x$ who would experience the same amount of confounding if attending the comparison school and therefore are homogeneous in $b_{G0|X}$. We have observed $Y_{G0.T1}^{(1)}$ of post-policy students in the comparison school and $Y_{G0.T0}^{(1)}$ of pre-policy students in the comparison school. Prognostic score model specifications will allow us to predict this pair of potential outcomes for all four groups of students. Within each homogeneous subpopulation, DID analysis is expected to generate an unbiased estimate of the policy effect of interest.

This strategy is an extension of the Peters-Belson method and the prognostic score method. Suppose that the outcome generating mechanism within each treatment group is based on a set of observed pretreatment characteristics *X*. One may fit a regression model in the comparison group and then apply the fitted model to each experimental unit. The prediction model generates, for each experimental unit, a predicted counterfactual outcome associated with the comparison condition (Belson, 1956; Peters, 1941). This predicted outcome has been named a "prognostic score" because it is a function of prognostic pretreatment covariates *X* (Hansen, 2008).

The *X*-*Y* relationships may change from the pre-policy year to the post-policy year in the comparison school due to the confounding of concurrent events. Hence for our purpose, we estimate a pair of prognostic scores, $\psi_0^{G0}(X)$ and $\psi_1^{G0}(X)$, defined as the predicted pre-policy and post-policy outcomes respectively of a student if assigned to the comparison school. To simplify the notation, henceforth we use ψ_0 and ψ_1 as a shorthand for $\psi_0^{G0}(X)$ and $\psi_1^{G0}(X)$, respectively. Let

$$\psi_0 = E\left[Y_{G0.T0}^{(1)}|X = x\right] = f(x, 0),$$
$$\psi_1 = E\left[Y_{G0.T1}^{(1)}|X = x\right] = f(x, 1).$$

Hansen (2008) has shown that the pretreatment covariates *X* become independent of the potential outcomes given the corresponding prognostic scores, that is,

$$Y_{G0.T0}^{(1)} \perp X|\psi_0,$$
$$Y_{G0.T1}^{(1)} \perp X|\psi_1.$$

When *X* includes all the outcome predictors, which is nearly possible in high-quality educational accountability data that contain repeated assessments of students over multiple years, the amount of confounding for those attending the comparison school conditioning on *X* can be represented as

$$b_{G0|X} = E\left[Y_{G0.T1}^{(1)}|G = 0, T = 1, X\right] - E\left[Y_{G0.T0}^{(1)}|G = 0, T = 0, X\right] = \psi_1 - \psi_0.$$

5.2. Identification Assumptions

Suppose that ψ_0 and ψ_1 are based on true models for $Y_{G0.T0}^{(1)}$ and $Y_{G0.T1}^{(1)}$ respectively. Then the following results hold:

$$Y_{G0.T0}^{(1)} \perp X|\psi_0, \psi_1;$$
$$Y_{G0.T1}^{(1)} \perp X|\psi_0, \psi_1.$$

We now propose an alternative form of Assumption (4):

$$E\left[Y_{G1.T1}^{(1)}|G = 1, T = 1, \psi_0, \psi_1\right] - E\left[Y_{G1.T0}^{(1)}|G = 1, T = 0, \psi_0, \psi_1\right]$$
$$= E\left[Y_{G0.T1}^{(1)}|G = 0, T = 1, \psi_0, \psi_1\right] - E\left[Y_{G0.T0}^{(1)}|G = 0, T = 0, \psi_0, \psi_1\right]. \quad (8)$$

Assumption (8) is considerably weaker than assumption (6) invoked by the propensity score-based methods as the former does not require $E\left[Y_{G1.T1}^{(1)}|G=1,T=1,\psi_0,\psi_1\right] = E\left[Y_{G0.T1}^{(1)}|G=0,T=1,\psi_0,\psi_1\right]$. Nor does it require $E\left[Y_{G1.T0}^{(1)}|G=1,T=0,\psi_0,\psi_1\right] = E\left[Y_{G0.T0}^{(1)}|G=0,T=0,\psi_0,\psi_1\right]$. Assumption (8) also implies the following:

(9a) $\psi_t = f(x,t)$, for $t = 0,1$ defines the function for the counterfactual outcome under the comparison condition at time *t* regardless of one's actual treatment group membership. This assumption is stronger than assumption (4) because it additionally requires that, if an experimental unit had counterfactually been assigned to the comparison school in a given time period *t*, the *x*-*y* relationship would have been the same as that of the comparison units with observed pretreatment characteristics *x*. Hence $\psi_1 - \psi_0 = f(x,1) - f(x,0)$ is the expected amount of confounding associated with concurrent events under the comparison condition for all units with observed pretreatment characteristics *x*. However, assumption (9a) is different from assumption (7a) invoked by the nonlinear CIC method that requires applying a single production function $h(u,1)$ to both $Y_{G1.T1}^{(1)}$ and $Y_{G0.T1}^{(1)}$ when $t=1$ and a single production function $h(u,0)$ to both $Y_{G1.T0}^{(1)}$ and $Y_{G0.T0}^{(1)}$ when $t=0$.

(9b) The support for the observed covariates *X* in the comparison school, denoted by $\mathbb{X}_0$, encompasses the support in the experimental school denoted by $\mathbb{X}_1$. This is similar to Athey and Imbens' (2006) assumption (7d) with regard to the unobservable *U*.

We prove in the Appendix that the DID estimator integrated over the distributions of the two prognostic scores is an unbiased estimate of $\delta_{G1.T0}$ under the above assumptions. As a side-note, if the *x*-*y* association does not change from the pre-policy year to the post-policy year in the comparison school, the two prognostic scores will differ only by a constant. In other words, the amount of confounding does not depend on X. Then the standard DID will apply.

5.3. Analytic Procedure

To implement, we specify a prognostic score model for the comparison school students in the pre-policy year. In parallel, we specify a second prognostic score model for the comparison school students in the post-policy year. We then apply these two models to all students in all four *G*-by-*T* combinations. We predict every student's prognostic score associated with the comparison condition in the pre-policy year and that associated with the comparison condition in the post-policy year. Hence every student has a pair of prognostic scores ψ_0 and ψ_1. Assuming that the confounding effect of concurrent events is the same across the comparison school and the experimental school for students who are homogeneous in the prognostic scores, we then conduct DID within cells jointly defined by ψ_0 and ψ_1. For example, we may divide the sample into three strata on the basis of ψ_0 and then subdivide each stratum into three on the basis of ψ_1. We may conduct a standard DID analysis within each of the nine cells and then pool the results to obtain an estimate of the policy effect. This procedure allows the DID estimate to differ across different levels of ψ_0 and ψ_1. Let D_s for $s = 1, \ldots, 9$ denote the nine cells. Through analyzing the following model

$$Y_i = \alpha_1 T_i + \beta_0 G_i + \beta_1 G_i T_i + \sum_{s=1}^{9} \lambda_{si} D_{si} + e_i.$$

We obtain β_1 as an estimate of the average policy effect. That is,

$$\beta_1 = \sum_{s=1}^{9} \{[E(Y|G=1,T=1,S=s) - E(Y|G=1,T=0,S=s)] - [E(Y|G=0,T=1,S=s) - E(Y|G=0,T=0,S=s)]\} pr(S=s).$$

It is easy to show that the above model is equivalent to a weighted analysis

$$Y_i = \alpha_0 + \alpha_1 T_i + \beta_0 G_i + \beta_1 G_i T_i + e_i, \quad (10)$$

where the weight ω can be computed as follows:

when $G = 1, T = 1, S = s$,

$$\omega = \frac{pr(G = 1|T = 1)}{pr(G = 1|T = 1, S = s)};$$

when $G = 0, T = 1, S = s$,

$$\omega = \frac{pr(G = 0|T = 1)}{pr(G = 0|T = 1, S = s)};$$

when $G = 1, T = 0, S = s$,

$$\omega = \frac{pr(G = 1|T = 0)}{pr(G = 1|T = 0, S = s)};$$

when $G = 0, T = 0, S = s$,

$$\omega = \frac{pr(G = 0|T = 0)}{pr(G = 0|T = 0, S = s)}.$$

To obtain an estimate of the policy effect on the untreated pre-policy students in the experimental school $\delta_{G1.T0}$, the DID estimate in each of the nine cells is to be weighted by the cell-specific proportion of pre-policy students in the experimental school. We can easily change the weight to the following:

when $G = 1, T = 1, S = s$,

$$\omega = \frac{pr(G = 1|T = 1)}{pr(G = 1|T = 1, S = s)} \times \frac{pr(G = 1|T = 0, S = s)}{pr(G = 1|T = 0)};$$

when $G = 0, T = 1, S = s$,

$$\omega = \frac{pr(G = 0|T = 1)}{pr(G = 0|T = 1, S = s)} \times \frac{pr(G = 1|T = 0, S = s)}{pr(G = 1|T = 0)};$$

when $G = 1, T = 0, S = s$,

$$\omega = 1;$$

when $G = 0, T = 0, S = s$,

$$\omega = \frac{pr(G = 0|T = 0)}{pr(G = 0|T = 0, S = s)} \times \frac{pr(G = 1|T = 0, S = s)}{pr(G = 1|T = 0)}.$$

As we will discuss in section 6.2, the weighted model (10) is relatively convenient to use in multi-cohort analysis.

Various semi-parametric and non-parametric strategies can be employed in specifying the prognostic score models. Issues related to model misspecifications are beyond the scope of the current paper. However, by allowing the models for ψ_0 and ψ_1 to be different functions of *X* under the comparison condition, *T* and *G* each take a fixed value in a prognostic score model. Hence there is no need to consider *T*-by-*X* interaction, *G*-by-*X* interaction, *T*-by-*G* interaction, and *T*-by-*G*-by-*X* three-way interactions in any given model. Finally, the prognostic score-basis DID method does not preclude covariance adjustment in the outcome model for further bias removal and precision improvement.

6. Discussion

A major challenge to DID analyses is that the confounding effect of concurrent events may vary by pretreatment covariates that are distributed differently across the experimental group and the comparison group. We have shown that the prognostic score-based DID strategy provides a new solution by allowing the average amount of confounding to differ between the experimental group and the comparison group. This new strategy invokes assumptions weaker than most of the existing DID methods. In particular, by using a prognostic score-based weighting adjustment, the outcome model is

non-parametric in nature and hence is exempt from strong model-based assumptions that make conventional DID analyses prone to bias.

In this section we first discuss the relative strengths and limitations of the prognostic score-based DID strategy in comparison with the existing alternative DID strategies. A particular focus is placed on comparing the identification assumptions across these different methods. We then extend the prognostic score-based DID strategy to multilevel multi-cohort data typically seen in education accountability systems.

6.1. Comparisons between Prognostic Score-Based DID and Other Existing DID Strategies

(a) Unlike the linear covariance adjusted DID models, the prognostic score-based DID strategy does not assume that, conditional on the observed pretreatment covariates *X*, the average treatment effect for the untreated in the experimental group is the same as that for the entire population should all units have been untreated. This advantage is shared by the propensity score-based DID methods and the nonlinear CIC models.

(b) Unlike linear covariance adjusted DID, prognostic score-based DID does not assume invariant *x*-*y* relationships across time.

(c) Prognostic score-based DID assumes that, if an experimental unit had counterfactually been assigned to the comparison condition in a given time period, the *x*-*y* relationship would have been the same as that of the comparison units with the same observed pretreatment characteristics. This differs from the assumption invoked by the nonlinear CIC method that requires applying a single production function to the outcomes of both the experimental group and the comparison group in a given time period.

(d) Unlike the nonlinear CIC models, the prognostic score-based DID models do not require strict monotonicity. Nor do they require that the outcome contain no measurement error. These advantages are shared by linear covariance adjusted DID and propensity score-based DID.

(e) While the nonlinear CIC models assume time invariance in the distribution of the unobserved *U* within the experimental group and the comparison group, this is not a requirement for all the other DID methods including prognostic score-based DID.

(f) A major difference between propensity score-based DID and prognostic score-based DID is that the latter does not require equating the pretreatment composition of the experimental group and the comparison group. The same advantage is shared by linear covariance adjusted DID and nonlinear CIC.

(g) Similar to propensity score-based DID, prognostic score-based DID emphasizes and verifies the common support between the experimental group and the comparison group with regard to observed covariates *X*, which effectively avoids unwarranted extrapolation. However, propensity score-based DID may suffer if some pretreatment covariates unrelated to the outcome lead to a shrinkage in the common support. The nonlinear CIC models make a similar assumption with regard to the unobserved covariates *U* that cannot be empirically verified.

(h) Similar to the propensity score-based DID, the prognostic score-based DID greatly reduces the dimensionality of covariates for adjustment, which is a major advantage over the linear covariance adjusted DID.

(i) Both prognostic score-based DID and DID with propensity score-based matching enable researchers to detect heterogeneity in the confounding effects of concurrent events as well as in the policy effect.

(j) A unique feature of the nonlinear CIC models is that it does not require explicit specification of the outcome model. All other DID strategies require explicit modeling that involves functional forms. The prognostic score-based DID models are no exception. To alleviate the impact of misspecifying the functional form of the model,

researchers may employ various semi-parametric or nonparametric approaches, a topic beyond the scope of this paper.

(k) Both linear covariance-adjusted DID and propensity score-based DID would suffer if the experimental group and the comparison group differ in the distribution of unobsered *U* and if the amount of confounding of concurrent events is a function of *U*. The same type of unobserved *U*, if independent of the observed covariates *X*, would also bias the prognostic score-based DID estimate of the policy effect. However, if the confounding does not depend on *U* or if the distribution of *U* is the same between the experimental group and the comparison group conditioning on *X*, then omitting *U* would not introduce bias in general. The special implication for the prognostic score-based DID method is that the estimated policy effect could possibly be unbiased even when the prognostic score models have low predictive power. The nonlinear CIC model allows the distribution of *U* to differ between the experimental group and the comparison group yet requires that the *U*-*Y* relationship be the same between these two groups. This assumption seems implausible because a change in the distribution of *U* will likely change the *U*-*Y* relationship.

Future research will compare the performance of prognostic score-based DID with that of other existing DID methods under different sets of assumptions through simulations. Sensitivity analysis may be developed to assess the amount of bias associated with a possible unobservable covariate.

6.2. Extension of the Prognostic Score-Based DID Strategy to Multilevel Multi-Cohort Data

We may extend the prognostic score-based DID strategy first to multiple experimental schools and multiple comparison schools enrolling one pre-policy cohort and one post-policy cohort. We then extend the method to time-series data of multiple cohorts of pre-policy and post-policy 9th graders.

In multilevel data, a student's potential outcome is a function of student-level pretreatment covariates *X* and school-level pretreatment covariates *W*. One may specify a pair of two-level prognostic score models with students at level 1 and schools at level 2 for the pre-policy and post-policy outcomes under the comparison condition. In theory, a student might have multiple prognostic scores depending on which comparison school the student might have counterfactually attended. We define the prognostic scores as a student's predicted outcome of attending a typical comparison school, which can be viewed as the average of the school-specific prognostic scores for the student. In accountability systems in which repeated assessments of student academic achievement have been equated vertically, one may model the growth trajectories of students as well.

The CPS data contain multiple cohorts of ninth graders both before and after the policy was introduced. Making full use of the available data may increase the statistical power. By modeling the systematic trend in outcome change over time in the absence of policy change, one may gain additional leverage in removing the confounding effect of concurrent events. More importantly, it becomes possible to investigate whether the policy effect was enhanced as its implementation became mature or whether the effect faded out over time if the reform lost its momentum after the initial period.

Suppose that the data include three pre-policy cohorts and three post-policy cohorts of ninth graders going through the same set of high schools. Let $t = -2, -1, 0$ denote the three pre-policy years 1996, 1995, and 1994, respectively. Let $t' = 1, 2, 3$ denote the three post-policy years 1997, 1998, and 1999, respectively. Past applications have relied heavily on model-based assumptions with regard to the temporal trend in the data in the absence of policy change. Applying the prognostic score-based DID strategy to multi-cohort data, we define the causal estimands non-parametrically and therefore do

not impose a linear time trend. Let $Y_{1t}^{(0)}$ denote the potential outcome of a pre-policy student attending an experimental school in the absence of the policy in pre-policy year *t*; let $Y_{1t'}^{(1)}$ for $t' = 1, 2, 3$ denote the student's three counterfactual outcomes in the three respective post-policy years. The average first-year policy effect on the math learning of pre-policy students attending experimental schools is defined as follows:

$$\sum_{t=-2}^{0} E\left(Y_{11}^{(1)} - Y_{1t}^{(0)}\right) pr(t|G = 1).$$

Here $pr(t|G = 1)$ is the proportion of pre-policy students in experimental schools entering the ninth grade in year *t*. The policy effect may depend on the maturity of implementation. This can be investigated by combining the results of pair-wise DID analyses. Each DID analysis contrasts one pre-policy cohort in year t with one post-policy cohort in year t' and is based on the corresponding pair of prognostic scores ψ_t and $\psi_{t'}$. Let $I(t' - t)$ for $t' = 1, 2, 3$ be the indicator for the subset of data used in the DID analysis for estimating the policy effect after t' years of implementation. Let $Z = 1$ denote the post-policy years and 0 for the pre-policy years. A weighted outcome model will be

$$Y_{ij} = \sum_{t'=1}^{3} I_{ij}(t' - t)\left(\alpha_{0t'} + \alpha_{1t'}Z_{ij} + \beta_{0t'}G_{ij} + \beta_{1t'}G_{ij}Z_{ij}\right) + u_j + e_{ij},$$
$$e_{ij} \sim N(0, \sigma^2), \quad u_j \sim N(0, \tau).$$

Here β_{11}, β_{12}, and β_{13} estimate the policy effects after one year, two years, and three years of implementation, respectively.

References

Abadie, A. (2005). Semiparametric difference-in-differences estimators. *Review of Economic Studies*. *72*, 1-19.

Athey, S. & Imbens, G. W. (2006). Identification and inference in nonlinear difference-in-differences models. *Econometrica*, *74*(2), 431-497.

Austin (2008). The performance of different propensity-score methods for estimating relative risks. *Journal of Clinical Epidemiology*, *61*(6), 537-545.

Barnow, B. S., Cain, G. G., & Goldberger, A. S. (1980). Issues in the analysis of selectivity bias. *Evaluation Studies*, *5*, 43-59.

Belson, W. A. (1956). A technique for studying the effects of a television broadcast. *Journal of the Royal Statistical Society*, *Series C* (Applied Statistics), *5*(3), 195-202.

Blundell, R., Costa Dias, M., Meghir, C. & Van Reenen, J. (2004), Evaluating the employment impact of a mandatory job search assistance program, *Journal of the European Economics Association*, *2*(4), 596-606.

Brookhart, A. M., Schneeweiss, S., Rothman, K., Glynn, R. J., Avorn, J., & Stürmer, T. (2006). Variable selection for propensity score models. *American Journal of Epidemiology*, *163*(12), 1149-1156.

Card, D. & Kruger, A. (1993). Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania. *American Economic Review*, *84*, 772-784.

Cerdá, M., Morenoff, J. D., Hansen, B. B., Tessari Hicks, K. J., Duque, L. F., Restrepo, A., & Diez-Roux, A. V. (2012). Reducing violence by transforming neighborhoods: A natural experiment in Medellín, Colombia. *American Journal of Epidemiology*. Advance access DOI: 10.1093/aje/kwr428.

Dynarski, S. (2003). Does aid matter? Measuring the effect of student aid on college attendance and completion. *The American Economic Review*, *93*, 279-288.

Fitzpatrick, M. D. (2008). Starting school at four: The effect of universal prekindergarten on children's academic achievement. *The B.E. Journal of Economic Analysis & Policy*, *8*(1) (Advances), Article 46.

Heckman, J. Ichimura, H., Smith, J., & Todd, P. (1998). Characterizing selection bias using experimental data. *Econometrica*, *66*(5), 1017-1098.

Heckman, J. Ichimura, H., & Todd, P. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *The Review of Economic Studies* Special Issue: Evaluation of Training and Other Social Programmes, *64*(4), 605-654.

Hansen, B. B. (2008). The prognostic analogue of the propensity score. *Biometrika*, *95*(2), 481-488.

Meyer, B. (1995). Natural and quasi-experiments in economics. *Journal of Business & Economic Statistics*, *13*, 151-161.

Peters, C. C. (1941). A method of matching groups for experiment with no loss of population. *Journal of Educational Research*, *34*(8), 606-612.

Appendix

Here we prove that the DID estimator integrated over the joint distribution of the two prognostic scores ψ_0 and ψ_1 is an unbiased estimate of $\delta_{G1.T0}$ under the identification assumptions (8), (9a), and (9b).

$$\iint_{\psi_1\psi_0} \left(D_{G1|\psi_1,\psi_0} - D_{G0|\psi_1,\psi_0}\right) f(\psi_0|\psi_1) f(\psi_1) d\psi_0 d\psi_1$$

$$= \iint_{\psi_1\psi_0} (\{E[Y|G=1,T=1,\psi_1,\psi_0] - E[Y|G=1,T=0,\psi_1,\psi_0]\} - \{E[Y|G=0,T=1,\psi_1,\psi_0] - E[Y|G=0,T=0,\psi_1,\psi_0]\}) f(\psi_0|\psi_1) f(\psi_1) d\psi_0 d\psi_1$$

$$= \iint_{\psi_1\psi_0} \left(\left\{E\left[Y^{(1)}_{G1.T1}|G=1,T=1,\psi_1,\psi_0\right] - E\left[Y^{(0)}_{G1.T0}|G=1,T=0,\psi_1,\psi_0\right]\right\} - \left\{E\left[Y^{(1)}_{G0.T1}|G=0,T=1,\psi_1,\psi_0\right] - E\left[Y^{(1)}_{G0.T0}|G=0,T=0,\psi_1,\psi_0\right]\right\}\right) f(\psi_0|\psi_1) f(\psi_1) d\psi_0 d\psi_1$$

$$= \iint_{\psi_1\psi_0} \left\{E\left[Y^{(1)}_{G1.T0}|G=1,T=1,\psi_1,\psi_0\right] - E\left[Y^{(0)}_{G1.T0}|G=1,T=0,\psi_1,\psi_0\right]\right\} f(\psi_0|\psi_1) f(\psi_1) d\psi_0 d\psi_1 + \iint_{\psi_1\psi_0} \left(\left\{E\left[Y^{(1)}_{G1.T1}|G=1,T=1,\psi_1,\psi_0\right] - E\left[Y^{(1)}_{G1.T0}|G=1,T=0,\psi_1,\psi_0\right]\right\} - \left\{E\left[Y^{(1)}_{G0.T1}|G=0,T=1,\psi_1,\psi_0\right] - E\left[Y^{(1)}_{G0.T0}|G=0,T=0,\psi_1,\psi_0\right]\right\}\right) f(\psi_0|\psi_1) f(\psi_1) d\psi_0 d\psi_1$$

$$= \delta_{G1.T0}.$$