DISCUSSION PAPER SERIES

高麗大學校 經濟硏究所

THE INSTITUTE OF ECONOMIC RESEARCH
KOREA UNIVERSITY

『Difference in Generalized-Differences with Panal Data : Effects of Moving from Private to Public School on Test Scores』

Myoung-jae Lee

Discussion Paper No. 07-21 (August 2007)

The Institute of Economic Research Korea University
Anam-dong, Sungbuk-ku,
Seoul, 136-701, Korea
Tel: (82-2) 3290-1632 Fax: (82-2) 928-4948

Difference in Generalized-Differences with Panel Data: Effects of Moving from Private to Public School on Test Scores

(August 25, 2007)

Myoung-jae Lee*

Dept.Economics, Korea University

Anam-dong, Sungbuk-gu

Seoul 136-701, South Korea

myoungjae@korea.ac.kr

phone: 82-2-3290-2229; fax: 82-2-926-3601

Difference in differences (DD) relies on the key identifying condition that the untreated response variable would have grown equally across the control and treatment groups; i.e., the 'time effects' across the groups are the same. This condition can be rewritten as the 'group effects' across the time points being the same, with which this paper generalizes DD to *difference in generalized-differences (DG).* DG is indexed by a parameter η, and includes DD as a special case when $\eta = 1$. This makes it possible to use DG as a sensitivity analysis for DD by trying values of η other than one. Going further from sensitivity analysis, one may desire to fix η. For this, we provide a way to get a benchmark value of η using a dynamic panel data model for the control group. An empirical analysis is provided for the *effects of moving from private to public school* on test scores. In the empirical analysis, (i) DD magnitude is fairly sensitive to changes in η around one, but its statistical significance is not, (ii) η is significantly smaller than one in math score (and possibly in science score), and (iii) DG yields a significant negative effect of about 3-5% for reading score, but the effects are ambiguous or insignificant for the other scores. η being less than 1 means that, had the movers to public school stayed, the score gap between the movers and stayers would have narrowed. That is, the move to public school is likely to have been involuntary.

Key Words: treatment effect, difference in differences, panel data, private-school effect.

* The author is grateful to the Editor, Associate Editor, two reviewers for their detailed comments which led to a substantial improvement of the paper. The author is also grateful to participants in various seminars at which the paper was presented.

1 Introduction

Difference in differences (DD) is one of the most popular—and often convincing—study design in finding the effects of a treatment. For instance, Bertrand et al. (2004) found as many as 92 papers in six economic journals over 1990-2000. Many DD references and examples can be found in Meyer (1995), Angrist and Krueger (1999), Heckman et al. (1999), Rosenbaum (2002), and Besley and Case (2004), just to name a few.

To briefly introduce DD, consider a treatment given at some time point between t_0 and t_1. There are two regions, $r = 0, 1$, and the treatment is given only to region 1, which makes region 1 the treatment group (*T group*) and region 2 the control group (*C group*). For an individual i with his/her response y_{it} at time t, $E(y_{it_1} - y_{it_0}|r = 0)$ includes only the time effect whereas $E(y_{it_1} - y_{it_0}|r = 1)$ includes both the time and treatment effects. Thus

$$DD \equiv E(y_{it_1} - y_{it_0}|r = 1) - E(y_{it_1} - y_{it_0}|r = 0)$$

identifies the desired effect by removing the time effect. The next section will show that the key DD identifying assumption is that, without the treatment, the time effects across the two regions are the same; that is, had the treatment been withheld in region 1 contrary to the fact, then both regions' response variables would have grown by the same degree. *Under this same time-effect condition, DD is known to identify the treatment effect on* $r = 1$ *at* $t = t_1$. From now on, we will assume iid across $i = 1, ..., N$ to often omit the subscript i.

The same time-effect condition, which allows the baseline response y_{t_0} to differ across the two groups, has been used without much challenge in the literature. But this is odd for the following reason. If the treatment is randomized, then y_{t_0} is balanced and there is no reason to use DD, as one can just compare y_{t_1} across the two groups. Thus DD is useful mainly for non-experimental data. The same time-effect condition, however, is a 'selection-on-observables' that, given the observables x, r is independent of the temporal change in the untreated response, which amounts to a randomization of r given x. In observational data, this randomization may not hold. It is thus imperative to examine in which direction the same time-effect condition may be violated and what to do about the problem.

One way to allow for the violation is *'generalized-difference in differences' (GD)* indexed by a single parameter γ:

$$GD_\gamma \equiv E(y_{t_1} - y_{t_0}|r = 1) - \gamma E(y_{t_1} - y_{t_0}|r = 0) \quad \Longrightarrow \; GD_1 = DD.$$

Surprisingly, it will be shown that GD_γ identifies the same treatment effect for any γ as DD does, but GD_γ invokes a different identification condition: the time effect in $r = 1$ is γ times the time effect in $r = 0$. This shows that DD sets $\gamma = 1$ a priori, for which little justification is offered in some DD applications. One thing that can be done with GD is a sensitivity analysis of DD: examine how GD_γ changes as γ changes around one. If GD_γ changes little despite γ changes much, then GD_γ is insensitive to γ, and thus using $GD_1 = DD$ would not matter much. Going further, it will be nice to find which identification condition—that is, which value of γ—is the right one. Unfortunately, GD does not seem to provide an useful answer to this question, but the following generalization of DD does.

Another way to embed DD is *'difference in generalized-differences' (DG)* indexed by η:

$$DG_\eta \equiv E(y_{t_1} - \eta y_{t_0} | r = 1) - E(y_{t_1} - \eta y_{t_0} | r = 0) \quad \Longrightarrow \; DG_1 = DD.$$

It will be shown that DG_η identifies the same treatment effect as DD and GD do, but DG_η invokes an identification condition different from those for DD and GD. We will discuss DG far more extensively than GD, because the parameter η may be found in a data-dependent and constructive way, whereas this does not seem to be the case for γ as noted already. Specifically, η will be identified as the coefficient of y_{t_0} in a dynamic panel data model.

Section 2 introduces notations and reviews DD. Section 3 examines DG in detail and discusses how to find η; Section 3 also briefly reviews GD. Section 4 presents an empirical analysis on the effects of moving from private to public school. In view of debates on whether or not charter, private, or Catholic schools have effects on academic achievements relative to public schools, this empirical analysis should be of interest on its own. Finally, Section 5 concludes.

2 Difference in Differences (DD)

Let $j = 0$ denote no treatment (or untreated state) and $j = 1$ denote the treatment given (or treated state). Also let y_{it}^j denote the 'potential' response at time t ($t = t_0, t_1$ with $t_0 < t_1$) when individual i received the treatment $j = 0, 1$ 'exogenously' at some time point just before t. Among the two potential responses (or outcomes) at any given time corresponding to the two potential treatments, only one outcome is observed while the other, called '*counter-factual*', is not. There are variables other than the treatment that affect the response variable; call them covariates (x_{it}) if observed, and error terms (u_{it}) if unobserved.

While we want to know the effects of the treatment on the response variable, if the covariates or the error terms take different values (or if they are unbalanced) across the T and C groups, then they can cause biases as well known. Let $x_i \equiv (x'_{it_0}, x'_{it_1})'$.

Define d_{it} as the treatment indicator variable, and

$$\begin{aligned} r_i &= 1 \text{ if person } i \text{ is in } r\text{egion 1, and 0 otherwise in } r\text{egion 0,} \\ \tau_t &= 1 \text{ if } t = t_1, \text{ and 0 otherwise if } t = t_0 \\ \Longrightarrow \quad & d_{it} = r_i \cdot \tau_t : \text{ being in region 1 at time } t_1 \text{ means treated.} \end{aligned}$$

Also define the 'observed response' $y_{it} = (1 - d_{it})y^0_{it} + d_{it}y^1_{it}$.

With x conditioned on, DD is

$$\begin{aligned} DD(x) &\equiv E(y_{t_1} - y_{t_0}|x, r = 1) - E(y_{t_1} - y_{t_0}|x, r = 0) \\ &= E(y^1_{t_1} - y^0_{t_0}|x, r = 1) - E(y^0_{t_1} - y^0_{t_0}|x, r = 0) \\ &= E(y^1_{t_1} - y^0_{t_0}|x, r = 1) - E(y^0_{t_1} - y^0_{t_0}|x, r = 1) \\ + \quad & E(y^0_{t_1} - y^0_{t_0}|x, r = 1) - E(y^0_{t_1} - y^0_{t_0}|x, r = 0), \end{aligned}$$

subtracting and adding $E(y^0_{t_1} - y^0_{t_0}|r = 1)$. *If the 'same time-effect' condition across the two regions holds* conditional on x, that is, if

$$E(y^0_{t_1} - y^0_{t_0}|x, r = 1) = E(y^0_{t_1} - y^0_{t_0}|x, r = 0), \qquad (\text{ID}_{DD})$$

then the second part of $DD(x)$ drops out, leaving

$$DD(x) = E(y^1_{t_1} - y^0_{t_0}|x, r = 1) - E(y^0_{t_1} - y^0_{t_0}|x, r = 1) = E(y^1_{t_1} - y^0_{t_1}|x, r = 1) :$$

DD identifies the effect for the treated region at the post-treatment period. Integrating out x in $DD(x)$ using the distribution $F(\cdot|r = 1)$ of $x|(r = 1)$ yields the 'marginal' DD:

$$DD = E\{\ E(y_{t_1} - y_{t_0}|x, r = 1) - E(y_{t_1} - y_{t_0}|x, r = 0)\ |r = 1\ \}.$$

Although we conditioned on both x_{t_0} and x_{t_1}, so long as ID_{DD} holds, conditioning on only x_{t_0} may be sufficient in some cases. Or the covariate x_t may well be time-constant. See Lee and Kang (2006) for more on DD identification conditions under various scenarios and data.

Since the treated response is observed only for region 1 at time t_1, unless some assumptions are imposed, it is natural that only the treatment effect for region 1 at time

t_1 is identified. Section 1 noted that $E(y_{t_1} - y_{t_0}|r = 0)$ includes the time effect, while $E(y_{t_1} - y_{t_0}|r = 1)$ includes the time effect and the treatment effect. To be precise, what is included in $E(y_{t_1} - y_{t_0}|r = 1)$ is the time effect and the '*interaction effect* between time t_1 and region 1 (the time-constant region effect is removed in $y_{t_1} - y_{t_0}$ for each region). Our empirical example later considers the effect of moving from private to public school. There the C group consists of those in private school at time both t_0 and t_1, whereas the T group is those in private school at t_0 and public school at t_1.

There may be unobserved confounders affecting (the mean of) the baseline response $y^0_{t_0}$ differently across $r = 0$ and $r = 1$. But so long as they affect $y^0_{t_1} - y^0_{t_0}$ in the same way across $r = 0$ and $r = 1$ such that ID_{DD} holds, the unobserved confounders do not cause a bias for DD. In words, ID_{DD} is that $y^0_{t_1} - y^0_{t_0}$ is mean-independent of the group indicator r given x; e.g., if y_t is $\ln(\text{GDP}_t$ per capita), then ID_{DD} requires both regions's untreated GDP mean growth rates to be the same conditional on x. But as already noted, ID_{DD} is a selection-on-observable assumption, which may not hold: as y_{t_0} may be unbalanced across the two groups, the untreated $y_{t_1} - y_{t_0}$ may be as well. Thus it is natural to think of in which direction ID_{DD} might be violated. This papers examines one such direction DG in detail and another such direction GD briefly. Certainly this paper is not the first one questioning ID_{DD}. For instance, Card and Sullivan (1988) noted that DD is not readily applicable to nonlinear models, and Meyer (1995) remarked that the plausibility of DD varies depending on which function of y is used as the response variable. Lalonde (1986) used DD while controlling for the lagged response, which may be taken as DG when the lagged response as a regressor is merged into the lagged response in DD.

3 Difference in Generalized-Differences (DG)

3.1 DG and Sensitivity Analysis for DD

With x conditioned on, DG is

$$\begin{aligned} DG_\eta(x) &= E(y_{t_1} - \eta y_{t_0}|x, r = 1) - E(y_{t_1} - \eta y_{t_0}|x, r = 0) \\ &= E(y^1_{t_1} - \eta y^0_{t_0}|x, r = 1) - E(y^0_{t_1} - \eta y^0_{t_0}|x, r = 0) \\ &= E(y^1_{t_1} - \eta y^0_{t_0}|x, r = 1) - E(y^0_{t_1} - \eta y^0_{t_0}|x, r = 1) \\ &+ \quad E(y^0_{t_1} - \eta y^0_{t_0}|x, r = 1) - E(y^0_{t_1} - \eta y^0_{t_0}|x, r = 0). \end{aligned}$$

By collecting the terms with η attached, it can be seen that $DG_\eta(x)$ is monotonic in η. If

$$E(y^0_{t_1} - \eta y^0_{t_0}|x, r = 1) = E(y^0_{t_1} - \eta y^0_{t_0}|x, r = 0), \qquad (\mathrm{ID}_{DG})$$

then

$$DG_\eta(x) = E(y^1_{t_1} - y^0_{t_1}|x, r = 1) :$$

regardless of the value of η, DG identifies the same treatment effect as DD does for the treated region at the post-treatment period. Integrating out x with $F(\cdot|r = 1)$ gives the *marginal DG that is the same as the marginal DD.* Clearly ID_{DG} includes ID_{DD} as a special case when $\eta = 1$, which leads to a sensitivity analysis for DD with DG by varying η. DD fixes $\eta = 1$ in ID_{DG} often without a due justification.

To better understand DG, rewrite ID_{DG} as

$$\begin{aligned} \{E(y^0_{t_1}|x, r &= 1) - E(y^0_{t_1}|x, r = 0)\} = \eta\{E(y^0_{t_0}|x, r = 1) - E(y^0_{t_0}|x, r = 0)\} \qquad (\mathrm{ID'}_{DG}) \\ &\iff \text{(region effect at time } t_1 \text{ given } x) = \eta \times \text{(region effect at time } t_0 \text{ given } x). \end{aligned}$$

This demonstrates that *DD removes the region effect assumed to be the same at time t_0 and t_1, whereas DG_η allows the region effect to differ between time t_0 and t_1.*

Since the left-hand side of ID_{DG} is counter-factual, ID_{DG} cannot be tested. In practice, one may estimate DG_η multiple times for, say, $\eta = 0.5$, 0.75, 1, 1.25, and 1.5 to compare the effects. Sensitivity in η means that the same region-effect assumption across the time points in DD is suspect. This kind of sensitivity analysis to relax assumptions of selection on observables is relatively new and can be seen in Rosenbaum (2002), Imbens (2003), Lee (2004), Altonji et al. (2005), and Lee et al. (2007). If DG_η does not vary much as η changes from 1, then DG_η (and thus DD) is insensitive to η and a wide range of η-values may be used freely. If DG_η is sensitive to η, however, then η has to be fixed, and the question is then how to choose the η value. This question will be addressed later.

As DG generalizes DD with η, not just the identification condition, but also the resulting estimator's asymptotic variance changes. The variance depends on the data generating model. The appendix illustrates the DG asymptotic variance using the models M_o and M_d to appear in the following subsections.

3.2 DG for Confounding Interactions

To better understand DG, suppose that the treatment has no effect such that the treated response equals the untreated response. Also suppose that the following model holds for the untreated response (x is ignored to simplify exposition):

$$\begin{aligned} y_{it} &= \alpha_t + \beta_t r_i + v_{it} \text{ where } \alpha_t \ (\beta_t) \text{ is time-varying intercept (slope)} \qquad (\mathrm{M}_o) \\ &\quad \text{and } v_{it} \text{ is a mean-zero error independent of } r_i. \end{aligned}$$

In this model, β_t is the time-t region effect. The DG parameter η is obtained from ID'$_{DG}$: assuming $\beta_{t_0} \neq 0$ ($\beta_{t_0} = 0$ is examined in the appendix),

$$\eta = \frac{\text{region effect at time } t_1}{\text{region effect at time } t_0} = \frac{\beta_{t_1}}{\beta_{t_0}} \implies \beta_{t_1} = \eta \beta_{t_0}.$$

Observe

$$\begin{aligned} &\beta_t = \beta_{t_0} + (\beta_{t_1} - \beta_{t_0})\tau_t = \beta_{t_0} + \beta_{t_0}(\eta - 1)\tau_t \quad \text{using } \beta_{t_1} = \beta_{t_0}\eta \\ \implies \quad &y_{it} = \alpha_t + \{\beta_{t_0} + \beta_{t_0}(\eta - 1)\tau_t\} r_i + v_{it} = \alpha_t + \beta_{t_0} r_i + (\eta - 1)\beta_{t_0}\tau_t r_i + v_{it}. \end{aligned}$$

This shows that, under no treatment effect, *DG allows for confounders affecting only region 1 at $t = t_1$*, and the effect of the confounders is the slope $(\eta - 1)\beta_{t_0}$ of the interaction term $\tau_t r_i$, while DD assumes away such variables with $\eta = 1$.

It is helpful to examine GD at this stage for model M_o. The time effect for region 0 is $\alpha_{t_1} - \alpha_{t_0}$ and the time effect for region 1 is $\alpha_{t_1} - \alpha_{t_0} + \beta_{t_1} - \beta_{t_0}$. Thus, assuming $\alpha_{t_1} \neq \alpha_{t_0}$ ($\alpha_{t_1} = \alpha_{t_0}$ is examined in the appendix),

$$\gamma = \frac{\text{time effect at region 1}}{\text{time effect at region 0}} = 1 + \frac{\beta_{t_1} - \beta_{t_0}}{\alpha_{t_1} - \alpha_{t_0}} \implies \beta_{t_1} - \beta_{t_0} = (\gamma - 1)(\alpha_{t_1} - \alpha_{t_0}).$$

The slope $\beta_{t_0} + (\beta_{t_1} - \beta_{t_0})\tau_t$ of r_i can be written as $\beta_{t_0} + (\gamma - 1)(\alpha_{t_1} - \alpha_{t_0})\tau_t$ to yield

$$y_{it} = \alpha_t + \{\beta_{t_0} + (\gamma - 1)(\alpha_{t_1} - \alpha_{t_0})\tau_t\} r_i + v_{it} = \alpha_t + \beta_{t_0} r_i + (\gamma - 1)(\alpha_{t_1} - \alpha_{t_0})\tau_t r_i + v_{it}.$$

Under no treatment effect, GD allows for confounders affecting only region 1 at $t = t_1$, and the effect of the confounders is the slope $(\gamma - 1)(\alpha_{t_1} - \alpha_{t_0})$ of the interaction term $\tau_t r_i$, while DD assumes away such variables with $\gamma = 1$.

To better understand DG, GD, and confounding interactions, examine the figure under no treatment effect. Line $\overline{ACF}$ is for control group, $\overline{BEI}$ is for treatment group, and there

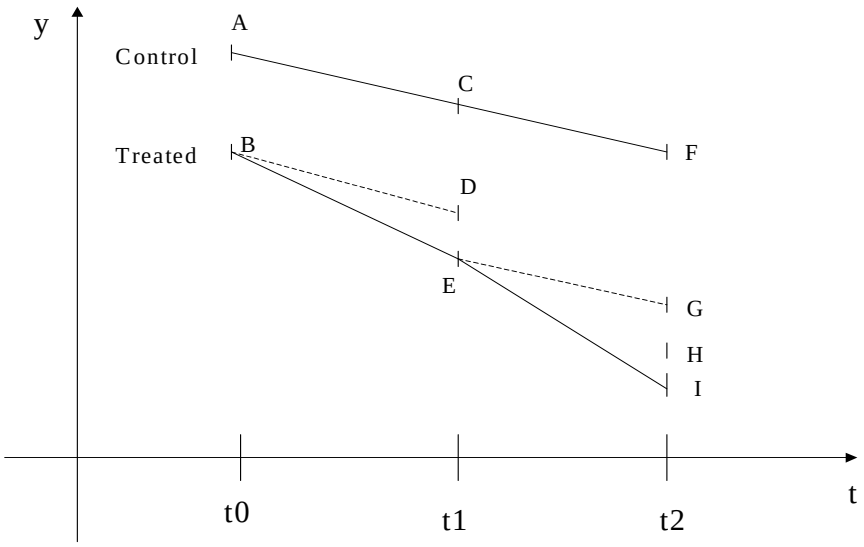

Figure 1: DG, GD and Confounding Interactions under No Effect

are three periods t_0, t_1, t_2 such that $t_2 - t_1 = t_1 - t_0 = 1$. Consider only t_0 and t_1 where t_1 is the post-treatment period. With *D-B=C-A*,

$$DD = E - B - (C - A) = (E - D) + (D - B) - (C - A) = E - D :$$

DD will take E-D as the treatment effect. But DG attributes E-D to the time-varying region effect, whereas GD attributes E-D to the region-varying time effect, i.e., the slope difference between $\overline{AC}$ and $\overline{BE}$. For two periods, DG and GD are two equivalent ways to account for the confounding interactions.

The figure also shows a way to find η (γ) and use the information if three periods are available; see Meyer (1995) and Shadish et al. (2002) among others for more on similar ideas. Suppose now t_0 and t_1 are pre-treatment periods while t_2 is a post-treatment period. In the figure,

$$\begin{aligned} \gamma &= \frac{E-B}{C-A} = \frac{E-D+D-B}{C-A} = \frac{E-D}{C-A} + 1; \\ \eta &= \frac{E-C}{B-A} = \frac{E-D+D-C}{B-A} = \frac{E-D}{B-A} + 1. \end{aligned}$$

The appendix shows that, if γ continues to hold for t_2, then the T group response at t_2 is

point H, which is $E + \gamma(F - C)$, and that, if η continues to hold for t_2, then the T group response at t_2 is point I, which is $F + \eta(E - C)$.

3.3 Finding a Benchmark DG Parameter

We proposed using DG as a sensitivity analysis for DD. While this is helpful, one may desire to find the DG parameter η, which is explored in this subsection. In a nutshell, we propose to identify η as the coefficient of the lagged response in a dynamic model for the control group $r = 0$. For this, we provide two justifications, one informal and the other formal.

An informal justification comes from the question why we ever wants to look at the difference $y_{t_1} - y_{t_0}$ instead of the level y_{t_1}. For our empirical example, why do we look at the test score *change* instead of the test *level*? If we have an experimental data, then y_{t_0} will be balanced across the two regions $r = 0, 1$, that is $E(y_{t_0}|r = 1) = E(y_{t_0}|r = 0)$, and we can just use $E(y_{t_1}|r = 1) - E(y_{t_1}|r = 0)$ for treatment effect. But for observational data, in general, $E(y_{t_0}|r = 1) \neq E(y_{t_0}|r = 0)$. This motivates DD, which reduces to $E(y_{t_1}|r = 1) - E(y_{t_1}|r = 0)$ for experimental data. The appearance of y_{t_0} in DD is the DD's way of controlling for y_{t_0} to avoid the problem $E(y_{t_0}|r = 1) \neq E(y_{t_0}|r = 0)$. That is, DD controls y_{t_0} in a simple way by setting $\eta = 1$ in $y_{t_1} - \eta y_{t_0}$, which does not require any estimation of η. But this simple solution does not work when $y_{t_1} - y_{t_0}$ still depends on y_{t_0}: the change in the test score can depend on the level of the baseline score, in which case the coefficient of y_{t_0} should be estimated. A simpler way of controlling y_{t_0} is using $y_{t_1} - \eta y_{t_0}$ as in DG when a dynamic model such as $y_{it_1} = \eta y_{it_0} + ...$ holds.

Turning to a formal justification, define

$$Y_i^0(\eta) \equiv y_{it_1}^0 - \eta y_{it_0}^0.$$

ID_{DG} assumes the mean-independence of $Y^0(\eta)$ from r given x, as the mean of $Y^0(\eta)$ is free of r given x. Strengthen this to the statistical independence of $Y^0(\eta)$ from r given x—not just the conditional mean but the entire conditional distribution of $Y^0(\eta)$ given x is free of r:

$$Y^0(\eta) \text{ are independent of } r \text{ given } x. \qquad (\text{ID}''_{DG})$$

In this case, Rosenbaum and Rubin (1983) show that $Y^0(\eta)$ is independent of r given only the scalar $P(r = 1|x) \equiv \pi(x)$—called the 'propensity score'. Thus,

$$E\{Y^0(\eta)|\pi(x), r = 1\} = E\{Y^0(\eta)|\pi(x), r = 0\}.$$

Also suppose that a single index assumption holds for the propensity score:

$$\pi(x) = \Phi(x'\xi)$$

where Φ is a strictly increasing distribution function and ξ is a parameter. Here $\pi(x)$ depends on x only through the index $x'\xi$.

Recall the no-effect model M_o but with $\beta_x' x_{it}$ augmented and the error term v_{it} decomposed as $v_{it} = \delta_i + u_{it}$ where δ_i is a time-constant error and u_{it} is a time-variant error. From the model, we get

$$\begin{aligned}
& y_{it_1} = \alpha_{t_1} + \beta_{t_1} r_i + \beta_x' x_{it_1} + \delta_i + u_{it_1} \quad \text{and} \quad y_{it_0} = \alpha_{t_0} + \beta_{t_0} r_i + \beta_x' x_{it_0} + \delta_i + u_{it_0} \\
\Longrightarrow \quad & Y_i^0(\beta_y) = y_{it_1} - \beta_y y_{it_0} \\
= \quad & \alpha_{t_1} - \beta_y \alpha_{t_0} + (\beta_{t_1} - \beta_y \beta_{t_0}) r_i + (\beta_x' x_{it_1} - \beta_y \beta_x' x_{it_0}) + (1 - \beta_y)\delta_i + (u_{it_1} - \beta_y u_{it_0}) \\
\Longrightarrow \quad & y_{it_1} = \beta_y y_{it_0} + (\alpha_{t_1} - \beta_y \alpha_{t_0}) + (\beta_{t_1} - \beta_y \beta_{t_0}) r_i \qquad (\mathrm{M}_d) \\
& + (\beta_x' x_{it_1} - \beta_y \beta_x' x_{it_0}) + (1 - \beta_y)\delta_i + (u_{it_1} - \beta_y u_{it_0}).
\end{aligned}$$

In M_d, $Y^0(\beta_y)$ becomes independent of r_i iff $\beta_y = \eta$. That is, under ID"$_{DG}$, η should be β_y, the slope coefficient of the lagged response in M_d. Hence η can be estimated from M_d, and we can use only the C group for the estimation because there is no selection problem under ID"$_{DG}$. When the treatment has an effect, β_{t_1} in the above non-dynamic y_{it_1} equation gets replaced by, say $\tilde{\beta}_{t_1}$, which will then appear in M_d; let $\tilde{\eta} \equiv \tilde{\beta}_{t_1}/\beta_{t_0}$. As DG_η uses η for both groups, DG will detect the effect. Compare this to using $\tilde{\eta}$ for T group and η for C group as in $E(y_{t_1} - \tilde{\eta} y_{t_0} | x, r = 1) - E(y_{t_1} - \eta y_{t_0} | x, r = 0)$, which would fail to detect the effect.

Although M_o (and the ensuing M_d) may look like a restrictive linear model, M_o is in fact a nonparametric saturated model because

$$E(y_{t_1}|r=1) = \alpha_{t_1} + \beta_{t_1}, \;\; E(y_{t_1}|r=0) = \alpha_{t_1}, \;\; E(y_{t_0}|r=1) = \alpha_{t_0} + \beta_{t_0}, \;\; E(y_{t_0}|r=0) = \alpha_{t_0}$$

which has four parameters and four identified means. The only restriction in M_o is the independence of $v_{it_1} - \eta v_{it_0}$ from r_i (given x_i), which is ID"$_{DG}$. When $\beta_x' x_{it}$ was introduced into M_o, another restriction gets imposed that $\beta_x' x_{it}$ enters the model only linearly. As it will become clear shortly, however, we can estimate η with a nonparametric method not subject to the linearity restriction.

To show our nonparametric estimation scheme under ID"$_{DG}$ and $P(r = 1|x) \equiv \pi(x) = \Phi(x'\xi)$, define

$$\varepsilon_{it_1} \equiv Y_i^0(\eta) - E\{Y_i^0(\eta) | \pi(x_i), r_i = 0\} \iff y_{it_1}^0 = \eta y_{it_0}^0 + E\{Y_i^0(\eta) | \Phi(x_i'\xi), r_i = 0\} + \varepsilon_{it_1}.$$

This is a semi-linear equation with a nonparametric component. By construction

$$
\begin{aligned}
& E(\varepsilon_{t_1}|\Phi(x'\xi), r=0)=0 \iff E(\varepsilon_{t_1}|x'\xi, r=0)=0 \\
\Longrightarrow \quad & E\{g(x'\xi)\varepsilon_{t_1}|r=0\}=0 \quad \text{for any (square-integrable) function } g(x'\xi) \\
\Longrightarrow \quad & E\{(\text{any linear function of } x)\cdot\varepsilon_{t_1}|r=0\}=0 \quad (\text{under } E|x|^2<\infty).
\end{aligned}
$$

Thus, on the subpopulation $r=0$, *a linear function of x is a valid instrument for y_{it_0}*.

To see that η is identified now, series-approximate $E\{Y^0(\eta)|\pi(x), r=0\}$ with

$$
\begin{aligned}
& E\{Y^0(\eta)|\pi(x), r=0\}=\zeta_0+\zeta_1\pi(x)+\zeta_2\pi(x)^2+,...,+\zeta_p\pi(x)^p \\
\Longrightarrow \quad & y_{it_1}=\eta y_{it_0}+\zeta_0+\zeta_1\Phi(x_i'\xi)+\zeta_2\Phi(x_i'\xi)^2+,...,+\zeta_p\Phi(x_i'\xi)^p+\varepsilon_{it_1}, \quad i\in C \quad (\mathrm{M}_n)
\end{aligned}
$$

where '$i \in C$' means that i belongs to the C group. This equation can be estimated by Instrumental Variable Estimator (IVE); an instrument is required only for y_{it_0}. A valid instrument for y_{it_0} is the fitted y_{it_0} in the Least Squares Estimator (LSE) of y_{it_0} on the baseline covariate x_{it_0} using the C group only. M_n avoids the linear parametric functional restriction $\beta_x' x_{it_1}$. M_n also avoids the lack of instrument problem in M_d: as both x_{t_1} and x_{t_0} appear in M_d, it is hard to find an instrument for y_{it_0}.

The propensity-score idea in Rosenbaum and Rubin (1983) was originally proposed as a way to avoid the dimension problem in estimating nonparametrically various expected values conditional on x. Because of this, instead of estimating $\pi(x)$ nonparametrically, $\pi(x)$ is almost always estimated parametrically under $\pi(x)=\Phi(x'\xi)$ in practice; nonparametric estimation of $\pi(x)$ would defeat the very motivation for propensity score. In the above derivation, propensity score is used also to identify η: the propensity score $\Phi(x'\xi)$ is necessarily *nonlinear* in x, which is essential for the validity of the *linear* (in x) instrument for the lagged response. If an arbitrary nonparametric function of x replaces $\zeta_1\Phi(x_i'\xi)+,...,+\zeta_p\Phi(x_i'\xi)^p$ in M_n, then IVE with a linear function of x as the instrument will fail.

When we plug the estimate $\hat{\eta}$ of η into the DG estimator, the resulting estimator becomes a two-stage procedure, and the estimation error $\hat{\eta}-\eta$ is likely to affect the asymptotic variance of the DG estimator. Deriving the two-stage asymptotic variance formally, however, goes beyond the scope of this paper. One approach to take $\hat{\eta}-\eta$ into account is bootstrap, resampling the original data multiple times with replacement. Because our problem is "smooth", the bootstrap inference is likely to work at least as well as the asymptotic inference. An alternative approach is to regard $\hat{\eta}$ as a fixed benchmark number rather than

an estimator for η, in which case $\hat{\eta} - \eta$ may be ignored and $\hat{\eta}$ may be treated as one of the fixed numbers used in sensitivity analysis. A practical middle ground for the two approaches is to obtain $\hat{\eta}$ and compare $DG_{\hat{\eta}}$ and its standard deviation (SD) to the sensitivity analysis DG estimates and SD's for η values close to $\hat{\eta}$. If not much difference is found among them, adopt the second approach of treating $\hat{\eta}$ as a fixed benchmark number; otherwise go for the bootstrap.

The above nonparametric idea uses the nonlinearity of $\Phi(\cdot)$ in an essential way, which may be undesirable because the realization of the nonlinearity depends on the support of $x'\xi$; reliance on the nonlinearity of $\Phi(\cdot)$ may make the identification of η fragile. For this, we can use a parametric dynamic panel data model as follows.

In M_d, the dynamic model was "artificial", as it was obtained as a transformation of a non-dynamic model. Suppose instead, when the treatment has no effect,

$$\begin{aligned} y_{it_1} &= \eta y_{it_0} + \alpha_{t_1} + \beta'_{t_1} x_{it_1} + \delta_i + u_{it_1} \ (y_{it_1} \text{ structural form (SF) free of } r_i \text{ for ID"}_{DG}) \ (\mathrm{M}_p) \\ y_{it_0} &= \alpha_{t_0} + \beta_{t_0} r_i + \beta'_{t_0} x_{it_0} + \beta_\delta \delta_i + u_{it_0} \ (y_{it_0} \text{ reduced form (RF) free of lagged responses}). \end{aligned}$$

We can imagine the y_{it_0} *SF* with a lagged response variable on the right-hand side. As this unavailable lagged response is substituted out successively, the y_{it_0} RF gets obtained, and r_i appears here because r_i appears in the very first period equation (the first year the child enters school). This process makes the parameters in the y_{it_0} RF different from those in the y_{it_1} SF as can be seen in M_p. Substitute the y_{it_0} RF into the y_{it_1} SF to get the y_{it_1} RF

$$y_{it_1} = (\eta\alpha_{t_0} + \alpha_{t_1}) + \eta\beta_{t_0} r_i + \eta\beta'_{t_0} x_{it_0} + \beta'_{t_1} x_{it_1} + (\eta\beta_\delta + 1)\delta_i + \eta u_{it_0} + u_{it_1}$$

which has the coefficient $\eta\beta_{t_0}$ for r_i: the region effect obtained from the y_{it_1}RF's is the η in the y_{it_1} SF. Apply IVE to M_p because y_{it_0} is an endogenous regressor in M_p. In our empirical analysis later, the school characteristics at t_0 is used as instruments for y_{it_0}, and the parametric version with M_p gives a more plausible result than the nonparametric version with M_n does.

In M_d, the non-dynamic model is a SF while the dynamic version M_d is the RF; in M_p, the opposite holds. In M_d, there is a time-varying region effect other than the treatment effect, and this confounding interaction is accounted for by the lagged response on the right-hand side of the y_{it_1} equation. In M_p, a true dynamic model is taken as a time-varying region effect, which is thus avoided by rightfully using the dynamic model. When there is a

treatment effect in model M_p, r_i will appear in M_p; e.g., r_i may enter M_p linearly or r_i may interact with y_{it_0} (i.e., the coefficient of y_{it_0} may vary across the regions). This effect will then picked up by DG_η.

In summary, DG using panel data can be implemented as follows:

1. Do probit of r on some covariates to obtain the propensity score $\Phi(x_i'\xi)$ for each i.

2. For each treated unit, select a matched control closest in terms of the propensity scores; the details of this matching is shown in the empirical section.

3. Obtain DG with $\eta = 1$ first, and then conduct sensitivity analysis with various η values around 1 for DG; the specific form of the DG estimator is provided also in the empirical section.

4. Find a benchmark η value by applying IVE to the nonparametric dynamic panel data model M_n or to the parametric dynamic panel data model M_p.

3.4 Generalized-Difference in Differences (GD)

Instead of $DG_\eta(x)$, consider GD conditional on x:

$$\begin{aligned} GD_\gamma(x) &\equiv E(y_{t_1} - y_{t_0}|x, r=1) \ - \gamma \cdot E(y_{t_1} - y_{t_0}|x, r=0) \\ &= E(y^1_{t_1} - y^0_{t_0}|x, r=1) - E(y^0_{t_1} - y^0_{t_0}|x, r=1) \\ + &\quad E(y^0_{t_1} - y^0_{t_0}|x, r=1) \ - \gamma \cdot E(y^0_{t_1} - y^0_{t_0}|x, r=0). \end{aligned}$$

Under the identifying condition

$$E(y^0_{t_1} - y^0_{t_0}|x, r=1) = \gamma \cdot E(y^0_{t_1} - y^0_{t_0}|x, r=0) \qquad (\mathrm{ID}_{GD})$$

$$\iff \quad \text{(time effect in region 1 given } x) = \gamma \times \text{(time effect in region 0 given } x),$$

$GD_\gamma(x)$ identifies the same treatment effect as $DG_\eta(x)$ and DD do:

$$GD_\gamma(x) = E(y^1_{t_1} - y^0_{t_1}|x, r=1).$$

The identifying condition is that, under no treatment, the $r = 1$ region would have grown by a degree "γ-proportional" to the growth of the $r = 0$ region. One may find ID_{GD} more agreeable than ID_{DG} because γ may be easier to interpret than η. But differently from DG, there seems to be no constructive data-dependent way of finding γ. Hence, GD may serve

only as a sensitivity analysis. Also, GD is equivalent to DG for two periods, which was shown graphically in the figure (and formally in the appendix). Since our empirical example has only two periods, GD will discussed only briefly after this subsection.

In ID_{GD}, γ appears multiplicatively, and one may wonder what if γ appears additively as in

$$E(y^0_{t_1} - y^0_{t_0}|x, r = 1) = E(y^0_{t_1} - y^0_{t_0}|x, r = 0) + \gamma'.$$

In this case, the appropriate GD, which becomes DD when $\gamma' = 0$, is

$$\begin{aligned} GD_{\gamma'}(x) &\equiv E(y_{t_1} - y_{t_0}|x, r = 1) - E(y_{t_1} - y_{t_0}|x, r = 0) - \gamma' \\ &= E(y^1_{t_1} - y^0_{t_0}|x, r = 1) - E(y^0_{t_1} - y^0_{t_0}|x, r = 1) \\ + &\quad E(y^0_{t_1} - y^0_{t_0}|x, r = 1) - E(y^0_{t_1} - y^0_{t_0}|x, r = 0) - \gamma' \end{aligned}$$

and the requisite identification condition is

$$E(y^0_{t_1} - y^0_{t_0}|x, r = 1) = E(y^0_{t_1} - y^0_{t_0}|x, r = 0) + \gamma'. \qquad (\text{ID'}_{GD})$$

For the "marginal version" (i.e., no x) of DG, equating the multiplicative and additive identification conditions and solving the equation for γ', we get $\gamma' = (\gamma - 1)E(y^0_{t_1} - y^0_{t_0}|r = 0)$: γ' is one-to-one to γ unless $E(y^0_{t_1} - y^0_{t_0}|r = 0) = 0$. Hence there is no loss in using the multiplicative form. If $E(y^0_{t_1} - y^0_{t_0}|r = 0) = 0$, however, then the additive GD should be used as ID_{GD} for the multiplicative GD is void. The additive GD is further discussed in the appendix.

The same type of question on multiplicative or additive form does not arise for DG, as the additive form will lead to the cancellation of η in DG. The question is related to 'transformation of variables' as noted in Meyer (1995). For example, multiplicative version may be suitable for $y = GDP$ while additive version may be better for $\ln(GDP)$.

4 Effects of Public School

This section provides an empirical analysis using DG. Although DG may be implemented with independent cross-section data over two periods, choosing η in the aforementioned way requires a dynamic model. Thus, panel data are better suited for DG. We use the National Educational Longitudinal Survey of 1988 (NELS88) released by The National Center for Education Statistics in the USA. *Our response variables are various logged test scores*, and

the treatment is being in public school at grade 10. Our *study population is those in private school at grade 8* (and responding to the survey at grade 8 and 10). That is, the *treatment of interest is the effects of moving from private to public school for those in private school at grade 8*. As for the response variable, in the literature of private-school effect, graduation rates, college entrance rates, or standardized test scores have been used as response variables. But graduation dummies and college entrance dummies are observed only after the treatment, and thus not suitable for DG; DG is feasible only with test scores.

The treatment variable is an interesting one, because it may provide a chance to see the reversed effect of moving from public to private school. That is, if private school has positive effects, then negative effects should be seen when the treatment is reversed. In general, a reversed treatment provides an opportunity to detect unobserved confounders; Lee (2005) shows some examples on this point. Viewing the switching effect from private to public school as the reverse effect of switching from public to private school is, however, subject to 'no state dependence' assumption. For instance, suppose students pick up bad study habits in public school while they do not pick up any good study habit in private school. Then the movers from public to private school as well as the stayers have bad study habits. But in the reverse case, the movers from private to public school will get bad study habits, while the stayers will not.

Effects of private school on academic achievements (measured by standardized test scores) have been hotly debated. The early literature using the High School and Beyond Survey is reviewed in Neal (1998) with the main finding that private schools—most private schools are Catholic schools in the USA—increases educational attainment, at least for urban minorities. To dissipate controversies about the self-selection issue of private school in observational data in the early literature, randomized studies were implemented. For a Milwaukee private school voucher experiment which ruled out religious private schools, Rouse (1998) found a positive effect of being selected into the program on math score, but mixed effects on reading; note that the effect is for being selected into the program, not for being in private school. But Krueger and Zhu (2004) found no significant effect of winning a voucher on test scores, using New York city voucher experimental data. Cullen et al. (2003) also reported little evidence of "sought-after" school effect, using Chicago public school choice program data (no voucher, but lottery was held for over-subscribed schools); again the effect is for winning the lottery, not for actually being in "sought-after" school. See Neal (2002) and the

references therein for more on private school effects (and voucher).

4.1 Sample and Methodology

Our initial sample had $N = 10099$, yielding the following transition table:

	10th grade private	10th grade public
8th grade private	1296	516
8th grade public	50	8237

Only 50 students moved from public to private schools, which are too few to assess the effects of private school. Instead, we selected the subsample with the eighth-grade school being private to find the effects of public school. This yielded $N = 1812 = 1296 + 516$. NELS88 is a large data set, but there are many item non-responses. Eliminating individuals with missings in the variables to be used, we came up with a final sample of $N = 1141$ with 27% moving from private to public school. Each individual has y_{t_0}, y_{t_1}, x, r.

Recall $Y_i^0(\eta) = y_{it_1}^0 - \eta y_{it_0}^0$ and define analogously

$$Y_i^1(\eta) \equiv y_{it_1}^1 - \eta y_{it_0}^0 \quad \text{and} \quad Y_i(\eta) \equiv (1 - r_i)Y_i^0(\eta) + r_i Y_i^1(\eta).$$

To simplify notations, '(η)' in $Y^j(\eta)$ will be omitted. To control x while making use of the mean-independence of Y^0 from r given $\pi(x)$ which is implied by ID_{DG}, define $f\{\pi(x)|r = 1\}$ as the conditional density of $\pi(x)|(r = 1)$ and observe

$$\begin{aligned}
& \int [E\{Y|\pi(x), r = 1\} - E\{Y|\pi(x), r = 0\}] f\{\pi(x)|r = 1\} d\{\pi(x)\} \\
= & \int [E\{Y^1|\pi(x), r = 1\} - E\{Y^0|\pi(x), r = 0\}] f\{\pi(x)|r = 1\} d\{\pi(x)\} \\
= & \int [E\{Y^1|\pi(x), r = 1\} - E\{Y^0|\pi(x), r = 1\}] f\{\pi(x)|r = 1\} d\{\pi(x)\} \\
= & \int [E\{y_{t_1}^1 - \eta y_{t_0}^0|\pi(x), r = 1\} - E\{y_{t_1}^0 - \eta y_{t_0}^0|\pi(x), r = 1\}] f\{\pi(x)|r = 1\} d\{\pi(x)\} \\
= & E(y_{t_1}^1 - y_{t_1}^0|r = 1).
\end{aligned}$$

Hence a sample analog for the first integral is a consistent estimator for $E(y_{t_1}^1 - y_{t_1}^0|r = 1)$, which is the effect at time t_1 on the treated (those in private school at grade 8 and in public school at grade 10). If we further assume that Y^1 is mean-independent of r given x, then the first integral with the pooled-sample density of $\pi(x)$ as the integrator equals $E(Y^1 - Y^0) = E(y_{t_1}^1 - y_{t_1}^0)$, which is the effect at time t_1 on the entire study population.

One sample analog for the first integral is, with N_1 denoting the T group size,

$$\frac{1}{N_1}\sum_{j\in T}(Y_j - Y_{mj})$$

where Y_{mj} is a matched control for the treated Y_j, obtained as the nearest control in the absolute propensity score difference, and '$j \in T$' means that subject j is in the T group. But even the nearest control may not be a good matched control if the propensity score difference is large. For this, we use 'caliper' c as follow. Denoting the probit estimator for $\pi(x)$ as $\pi_N(x)$ and defining $w_j \equiv \min_{i\in C}|\pi_N(x_j) - \pi_N(x_i)|$, we use only those treated with $w_j < c$:

$$\frac{1}{\sum_{j\in T} 1[w_j < c]}\sum_{j\in T}(Y_j - Y_{mj})1[w_j < c] = \frac{1}{\sum_{j\in T} 1[w_j < c]}\sum_{j\in T: w_j < c}(Y_j - Y_{mj})$$

where $1[A] = 1$ if A holds and 0 otherwise. For the standard deviation of this matched-pair sample average, we use the square root of

$$\frac{1}{\{\sum_{j\in T} 1[w_j < c]\}^2}\sum_{j\in T: w_j < c}\{(Y_j - Y_{mj}) - \overline{Y_j - Y_{mj}}\}^2.$$

In the pair matching, we will maintain the same "control reservoir" for each treated unit; that is, a control may be used for multiple treated subjects.

4.2 Preliminary Analysis

Table 1 presents DD without controlling for x where SD and t-values (tv; the estimate divided by its SD) are shown in parentheses. The T group shows a significant decline of 2% only in reading score, whereas the C group shows a significant decline in history score. Overall, other than for reading score, the C group did worse than the T group. As the result, contrary to our intuition, the final DD column shows positive effects of public school other than for reading score. The only significant positive effect of public school is in history score, whereas a nearly significant decline is seen for reading score. But Table 1 ignores differences in x, and as such, it should be taken only as preliminary.

TABLE 1: DD without Controlling Covariates

Scores	T group $\bar{y}_{t_1} - \bar{y}_{t_0}$ (sd,tv)	C group $\bar{y}_{t_1} - \bar{y}_{t_0}$ (sd,tv)	DD (sd,tv)
History	0.00058 (0.0072, 0.081)	-0.018 (0.0046, -3.80)	0.018 (0.0086, 2.11)
Math.	0.0050 (0.0053, 0.95)	0.0038 (0.0035, 1.07)	0.0012 (0.0063, 0.19)
Reading	-0.020 (0.0077, -2.66)	-0.0055 (0.0044, -1.25)	-0.015 (0.0088, -1.69)
Science	0.0027 (0.0085, 0.32)	-0.0021 (0.0050, -0.42)	0.0048 (0.0099, 0.49)

TABLE 2: Mean and SD of Regressors in the Initial and Final Samples

Regressors	Mean (Initial), SD (Initial)	Regressors	Mean (Initial), SD (Initial)
Gender	0.496 (0.47), 0.500 (0.50)	Parents Christian	0.628 (0.50), 0.483 (0.50)
Black	0.041 (0.088), 0.199 (0.28)	Parents married	0.904 (0.76), 0.295 (0.43)
Hispanic	0.054 (0.12), 0.227 (0.32)	# Siblings	1.916 (2.70), 1.35 (6.9)
Other races	0.067 (0.10), 0.251 (0.30)	8 Sch enrol 2	0.269 (0.27), 0.444 (0.44)
Dad edu2	0.181 (0.17), 0.386 (0.38)	8 Sch enrol 3	0.146 (0.21), 0.354 (0.41)
Dad edu3	0.255 (0.14), 0.436 (0.35)	8 Sch enrol 4	0.052 (0.12), 0.222 (0.33)
Dad edu4	0.318 (0.13), 0.466 (0.34)	8 Sch enrol 5	0.029 (0.15), 0.168 (0.35)
Mom edu2	0.207 (0.192), 0.405 (0.39)	10 Sch enrol 2	0.263 (0.23), 0.440 (0.42)
Mom edu3	0.252 (0.137), 0.434 (0.34)	10 Sch enrol 3	0.199 (0.28), 0.399 (0.45)
Mom edu4	0.201 (0.091), 0.401 (0.29)	10 Sch enrol 4	0.086 (0.24), 0.280 (0.43)
Mom part-time	0.206 (0.15), 0.405 (0.36)	8 Sch day 1	0.349 (0.24), 0.477 (0.42)
Mom full-time	0.533 (0.55), 0.499 (0.50)	8 Sch day 2	0.561 (0.63), 0.496 (0.48)
Mom white-collar	0.017 (0.024), 0.128 (0.15)	10 Sch day 1	0.280 (0.16), 0.449 (0.36)
Income2	0.083 (0.17), 0.276 (0.37)	10 Sch day 2	0.629 (0.57), 0.483 (0.50)
Income3	0.154 (0.18), 0.361 (0.38)	8 Sch suburban	0.401 (0.43), 0.490 (0.50)
Income4	0.231 (0.20), 0.422 (0.40)	8 Sch urban	0.513 (0.24), 0.500 (0.43)
Income5	0.294 (0.17), 0.456 (0.37)	10 Sch suburban	0.300 (0.40), 0.458 (0.49)
Income6	0.202 (0.049), 0.402 (0.22)	10 Sch urban	0.592 (0.27), 0.492 (0.45)
		8 Sch $\frac{students}{teachers}$	18.7 (18.6), 6.41 (10)

Table 2 lists the mean and SD of the regressors used in getting the probit estimator $\hat{\xi}_N$ for the propensity score $\pi_N(x) = \Phi(x'\hat{\xi}_N)$. The mean and SD in $(\cdot)$ are for the initial sample before the individuals with missings were eliminated. In the left half of Table 2, 'Dad edu#' shows the father's education level: the omitted base-case is high school graduation or below, 2 is for above high school but below college degree, 3 is for college degree, and 4 is for above college. 'Mom edu#' is defined in the same way. Mom part-time, full-time, and white collar show the mother's work status. Income 2-6 show the household income categories: the omitted base case is for below \$15,000, 2 for \$15,000-24,999, 3 for \$25,000-34,999, 4 for \$35,000-49,999, 5 for \$50,000-99,999, and 6 for \$100,000 or above. '# Siblings' is top-coded at 6 with its SD being 1.35. These household characteristics are for grade 8, as they are not

available for grade 10.

Turning to time-variant school characteristics in the right half of Table 2, '8 Sch enrol #' is the number of enrolled students in the grade 8th school: the omitted base case is 399 or below, 2 for 400-599, 3 for 600-799, 4 for 800-999, and 5 for 1000 or above. '10 Sch enrol #' is for the number of enrolled students in the grade 10th school: the omitted base case is 599 or below, 2 for 600-999, 3 for 1000-1599, and 4 for 1600 or above. '8 Sch day 1' is the number of school days in the grade 8th school being 130-175, and '8 Sch day 2' is for 176-180; the omitted base case is for 181 or above. '10 Sch day 1' and '10 Sch day 2' are defined in the same way. The SD of the students/teachers ratio is 6.41.

Comparing the means of the final sample to those of the initial sample, the final sample is biased to the white, more-educated, and richer. Also the households in the final sample are more Christian, smaller, and more urban. In most cases, the SD of the final sample is either smaller or similar to the SD of the initial sample, but in some cases such as '#Siblings' and 'grade 8 students/teacher ratio', the SD declined much in the final sample. The two reasons for these differences between the initial and final samples are the sample selection criterion of being in private school at grade 8 and the elimination of non-response subjects. The first reason is not a concern, for we declared that our study population is those in private school at grade 8, but the second reason is a concern. Our empirical findings will be 'externally valid' only if there is no selection problem due to non-responses.

The pseudo-R^2 defined for probit as in Mckelvey and Zavoina (1975)

$$\frac{\hat{\xi}' N^{-1} \sum_i (x_i - \bar{x})(x_i - \bar{x})' \hat{\xi}}{\hat{\xi}' N^{-1} \sum_i (x_i - \bar{x})(x_i - \bar{x})' \hat{\xi} + 1}$$

is 0.871, which is very high for micro-data. This pseudo R^2 can be interpreted as the usual R^2 for linear models, because the pseudo-R^2 becomes the same as the usual linear model R^2 when the numerator and denominator are multiplied by the error term SD. For this reason, the pseudo R^2 was recommended by Veall and Zimmermann (1996) among many other pseudo R^2's. The most explanatory power in the probit comes from the school characteristic variables. Using only those and excluding the family and individual characteristic variables still yield the pseudo R^2 0.859, which is only negligibly smaller than 0.871.

For our purpose, probit itself is not of the main interest; only the predicted value $\Phi(x'\hat{\xi}_N)$ is. Even if the probit model is misspecified, so long as the predicted value is close to that of the correct model, that will be enough for our purpose. This is why the probit estimates

were omitted from Table 2. The high pseudo-R^2 suggests that one be less concerned about unobserved confounders in our public school effect analysis. Of course, it is still possible that the eventual treatment effects, if they are not large enough, are explained away by the small presence of the confounders.

4.3 Main Results

TABLE 3: Baseline (Grade 8) Group Means and Differences

Regressors	T, C groups	sd (tv) for Dif.	Regressors	T, C groups	sd (tv) for Dif.
Gender	0.487, 0.499	0.033 (-0.37)	Parents Christian	0.753, 0.581	0.030 (5.75)
Black	0.035, 0.043	0.013 (-0.65)	Parents married	0.888, 0.910	0.020 (-1.06)
Hispanic	0.042, 0.059	0.014 (-1.25)	# Siblings	1.962, 1.899	0.088 (0.71)
Other races	0.051, 0.074	0.015 (-1.44)			
Dad edu2	0.179, 0.182	0.026 (-0.10)	8 Sch enrol 2	0.167, 0.308	0.027 (-5.31)
Dad edu3	0.215, 0.270	0.028 (-1.99)	8 Sch enrol 3	0.125, 0.154	0.023 (-1.30)
Dad edu4	0.202, 0.362	0.028 (-5.67)	8 Sch enrol 4	0.029, 0.060	0.013 (-2.50)
Mom edu2	0.196, 0.211	0.027 (-0.59)	8 Sch enrol 5	0.006, 0.037	0.0080 (-3.87)
Mom edu3	0.170, 0.282	0.026 (-4.25)	8 Sch day 1	0.240, 0.390	0.030 (-5.05)
Mom edu4	0.144, 0.222	0.025 (-3.16)	8 Sch day 2	0.606, 0.544	0.033 (1.89)
Mom part-time	0.231, 0.197	0.028 (1.24)	8 Sch suburban	0.420, 0.393	0.033 (0.81)
Mom full-time	0.538, 0.531	0.033 (0.23)	8 Sch urban	0.429, 0.544	0.033 (-3.47)
Mom white-collar	0.016, 0.017	0.0084 (-0.10)	8 Sch $\frac{students}{tet_0chers}$	21.5, 17.7	0.368 (10.2)
Income2	0.112, 0.072	0.020 (1.99)			
Income3	0.212, 0.133	0.026 (3.03)	History 8	8.60, 8.64	0.011 (-3.54)
Income4	0.292, 0.209	0.029 (2.82)	Math 8	8.59, 8.64	0.012 (-4.18)
Income5	0.247, 0.311	0.029 (-2.20)	Reading 8	8.63, 8.64	0.011 (-1.19)
Income6	0.074, 0.251	0.021 (-8.38)	Science 8	8.60, 8.63	0.012 (-2.56)

Table 3 compares the T and C groups at the baseline (grade 8). The T group has the lower parental education, less income, and is more Catholic. The T group grade-8 private school is smaller in size and more rural, and has longer school days and higher students/teacher ratio. Also the T group has 3-5% lower scores. Overall, the T group differs from the C group in many components of x, and it is certainly possible that the two groups also differ

in unobserved variables, casting doubt on the selection-on-observable assumption ID_{DD}.

Suppose $\eta > 1$, which means that, according to ID'$_{DG}$, the score gap would have expanded at time t_1, had the movers stayed contrary to the fact. Given the general perception that public school does the worse job in educating children, moving to public school in this case would lead to even worse test scores. What seems more logical is thus $\eta < 1$: the gap would have narrowed, had the movers stayed. That is, the move must have been involuntary, perhaps due to unavoidable external circumstances such as a sudden loss of family income. This point and the fact that the coefficient of the lagged response tends to be below one give higher weights to the estimates with $\eta \leq 1$ than to the estimates with $\eta > 1$.

TABLE 4: Public-School Effects on Test Scores with DG (sd,tv)

Caliper:	0.03	0.005	0.03	0.005
	History		Reading	
$\eta = 1.5$	0.024 (0.017, 1.45)	-0.001 (0.020, -0.06)	-0.047 (0.016, -2.99)	-0.056 (0.020, -2.87)
$\eta = 1.25$	0.016 (0.014, 1.14)	-0.001 (0.017, -0.05)	-0.045 (0.013, -3.48)	-0.048 (0.016, -2.91)
$\eta = 1$ (DD)	0.008 (0.012, 0.64)	0.000 (0.016, -0.03)	-0.043 (0.011, -3.94)	-0.040 (0.014, -2.80)
$\eta = 0.75$	0.000 (0.011, -0.04)	0.000 (0.015, 0.003)	-0.042 (0.010, -4.08)	-0.031 (0.013, -2.38)
$\eta = 0.5$	-0.009 (0.012, -0.73)	0.000 (0.015, 0.03)	-0.040 (0.011, -3.69)	-0.023 (0.014, -1.67)
	Mathematics		Science	
$\eta = 1.5$	-0.006 (0.013, -0.49)	-0.020 (0.015, -1.34)	0.023 (0.019, 1.20)	-0.021 (0.023, -0.92)
$\eta = 1.25$	-0.005 (0.010, -0.53)	-0.010 (0.012, -0.81)	0.020 (0.016, 1.24)	-0.012 (0.019, -0.67)
$\eta = 1$ (DD)	-0.004 (0.008, -0.51)	0.001 (0.010, 0.10)	0.016 (0.013, 1.24)	-0.004 (0.015, -0.25)
$\eta = 0.75$	-0.003 (0.009, -0.38)	0.012 (0.010, 1.13)	0.012 (0.011, 1.11)	0.005 (0.013, 0.36)
$\eta = 0.5$	-0.002 (0.011, -0.22)	0.022 (0.013, 1.79)	0.008 (0.011, 0.77)	0.013 (0.013, 1.02)

Table 4 shows the effect estimates of moving from private to public school. Two sets of estimates are shown for caliper $c = 0.03$ and 0.005, and for each caliper, five estimates are presented for $\eta = 1.5$, 1.25, 1, 0.75, and 0.5. For $c = 0.03$ and 0.005, the proportions of the T group subjects who found a matched control are 0.71 and 0.49, respectively. Before the matching, the average propensity score difference between the two groups was 0.644, which is drastically reduced after matching to 0.002 and 0.000 for $c = 0.03$ and 0.005, respectively. In the remaining rows, the mean effects (sd, tv) are shown for each test score and each value

of η.

Table 4 shows that the history-score effect is too variable or too small to conclude anything; math score seems to have gone down, but falls short of being significant; reading score has significantly decreased by 3-5%; and the effect on science score might have been positive, but all numbers there are insignificant. Overall, the DD effect magnitudes are sensitive as η changes around one except for reading score, but their statistical significances are not. The appendix shows that, for the error term $v_{it} = \delta_i + u_{it}$ with $V(\delta_i) = \sigma_\delta^2$ ("inter-personal variance") and $V(u_{it}^2) = \sigma_u^2$ ("intra-personal" variance) where $COR(\delta_i, u_{it}) = 0\ \forall t$ and u_{it_1} and u_{it_0} are iid, the DG estimator variance is the smallest at $\eta = 0$ if $\sigma_\delta^2 = 0$ and at $\eta = 1$ if $\sigma_u^2 = 0$ (in general, the minimizing η falls in between 0 and 1). Judging from the SD's, math score seems to have the case $\sigma_u^2 \simeq 0$ (i.e., $v_{it} \simeq \delta_i$), and in the other scores, the smallest SD is realized at η strictly between 0 and 1.

TABLE 5: DG Effect with LSE and IVE for Nonparametric Model M_n

Score		η, SD (ts for $\eta = 1$)	DG (sd,tv) for $c = 0.03$	DG (sd,tv) for $c = 0.005$
History	LSE	0.606, 3.53 (-0.11)	-0.005 (0.011, -0.45)	0.000 (0.015, 0.02)
	IVE	*1.009, 0.083 (0.11)*	*0.008 (0.012, 0.66)*	*0.000 (0.016, 0.03)*
Math	LSE	0.702, 3.06 (-0.10)	-0.003 (0.009, -0.35)	0.014 (0.011, 1.30)
	IVE	*0.862, 0.052 (-2.63)*	*-0.004 (0.008, -0.45)*	*0.007 (0.010, 0.70)*
Reading	LSE	0.628, 3.24 (-0.12)	-0.041 (0.010, -3.95)	-0.027 (0.013, -2.05)
	IVE	*0.929, 0.079 (-0.89)*	*-0.043 (0.011, -4.03)*	*-0.037 (0.014, -2.71)*
Science	LSE	0.590, 4.15 (-0.10)	0.010 (0.011, 0.91)	0.010 (0.013, 0.80)
	IVE	*0.984, 0.072 (-0.22)*	*0.016 (0.013, 1.24)*	*-0.003 (0.015, -0.22)*

Turning to finding a benchmark value for η, the left half of Table 5 presents LSE and IVE applied to M_n using the C group only. The unreported R^2's for the LSE hover around 0.5; also unreported are the t-values for y_{t_0} which are all highly significant The LSE's for η are all downward biased relative to the IVE's. 'ts for $\eta = 1$' is the test statistic value for the $H_0 : \eta = 1$, which is rejected only for math score. The right half of Table 5 shows the effect estimates corresponding to the estimated η values. All effects are either small or insignificant except those for reading score, which are significantly negative around -4%. Hence DD seems reasonable in all but reading score. Given the general perception that math

and science are closely linked, this discrepancy between math η and science η is curious. For each test score, we can find a bound for IVE $\hat{\eta}$; e.g., in history score, $\hat{\eta} \simeq 1$ which is bounded by $[0.75, 1.25]$. In Table 4, over the range $[0.75, 1.25]$ for history score, there is no reversal of any qualitative finding. Thus we did not go further for an attempt to take $\hat{\eta} - \eta$ into account for SD. Essentially the same thing can be said for the other scores.

Table 6 which is analogous to Table 5 shows the results of LSE and IVE for the parametric dynamic panel data model M_p. The y_{it_1} equation is linear in y_{it_0} and all covariates except the time-t_0 school characteristics that are used as the instruments for y_{it_0}. Table 6 results are not much different from those in Table 5 except the IVE for science score, which has $\eta = 0.693$ with $H_0 : \eta = 1$ rejected. This is in contrast to Table 5 where the IVE for science score has $\eta = 0.984$. This finding in Table 6 is more convincing than the corresponding number in Table 5, because science score is likely to behave similarly to math score although the resulting DG effect is small and insignificant in both tables for science score. The M_n identification feature relying on the nonlinearity of $\Phi(\cdot)$ might have hindered the IVE for M_n.

TABLE 6: DG Effect with LSE and IVE for Parametric Model M_p

Score		η, SD (ts for $\eta = 1$)	DG (sd,tv) for $c = 0.03$	DG (sd,tv) for $c = 0.005$
History	LSE	0.549, 3.72 (-0.12)	-0.007 (0.012, -0.61)	0.000 (0.015, 0.03)
	IVE	*0.846, 0.14 (-1.09)*	*0.003 (0.012, 0.24)*	*0.000 (0.015, 0.01)*
Math	LSE	0.673, 3.04 (-0.11)	-0.003 (0.009, -0.33)	0.015 (0.011, 1.39)
	IVE	*0.865, 0.065 (-2.09)*	*-0.004 (0.008, -0.46)*	*0.007 (0.010, 0.68)*
Reading	LSE	0.597, 3.48 (-0.12)	-0.041 (0.010, -3.90)	-0.026 (0.013, -1.96)
	IVE	*0.987, 0.15 (-0.09)*	*-0.043 (0.011, -3.96)*	*-0.039 (0.014, -2.78)*
Science	LSE	0.519, 3.78 (-0.13)	0.009 (0.011, 0.80)	0.013 (0.013, 0.98)
	IVE	*0.693, 0.12 (-2.66)*	*0.011 (0.011, 1.05)*	*0.007 (0.013, 0.51)*

Overall, using DG, we conclude a significantly negative effect of 3-5% for reading score. For the other scores, the effects are ambiguous or insignificant, taking positive as well as negative signs. We also find that DD exaggerates the negative effect on math (and science) score, as the estimated η value is significantly less than 1. But the DG and DD estimate difference for math (and science) score is negligibly small. These findings differ somewhat from those in Table 1 where no covariate was controlled.

5 Conclusions

In this paper, we proposed 'difference in generalized-differences' (DG) that is indexed by a single parameter η. DG includes the popular study design 'difference in differences' (DD) as a special case when $\eta = 1$. The treatment effect identified by DG is the same as that identified by DD: the mean effect for the treated in the post-treatment era. There is one true, but unknown, value of η, which does not have to be unity as DD specifies. Letting $\eta \neq 1$, DG generalizes DD by allowing for different region (group) effects across time points, and DG provides a sensitivity check on DD.

An empirical analysis for the effects of moving from private to public school was provided, in which the sensitivity analyses were conducted, and how to choose η was illustrated using a dynamic model for the control group. The empirical findings are: (i) η is significantly lower than one for math score (and possibly for science score), and η is almost one for the other scores; DD exaggerates the negative effect on math (and science) score although this exaggeration hardly biased the DD estimate, (ii) there is a significant negative effect of 3-5% on reading score, and (iii) the effects on the other scores were ambiguous or insignificant. As it turned out, the DG findings differed little from DD in our particular empirical analysis. But there is no reason to believe that this will also hold for other empirical cases.

DG provides an opportunity to take a second look at the critical DD identification condition—a selection-on-observable assumption—which has been somehow accepted in the literature without much scrutiny. Relaxing this condition with DG matters greatly, as DD is widely applied to observational data to find treatment effects.

APPENDIX

A.1 Case of $\beta_{t_0} = 0$ or $\alpha_{t_1} = \alpha_{t_0}$.

If $\beta_{t_0} = 0$ in DG, then $\beta_t = \beta_{t_1}\tau_t$ and thus $y_{it} = \alpha_t + \beta_{t_1}\tau_t r_i + v_{it}$. ID_{DG} requires

$$\alpha_{t_1} + \beta_{t_1} - \eta\alpha_{t_0} = \alpha_{t_1} - \eta\alpha_{t_0} \implies \beta_{t_1} = 0 \ \forall\eta\text{: } \eta \text{ is not identified.}$$

But $\beta_{t_0} = \beta_{t_1} = 0$ is an uninteresting case, because the baseline mean responses are balanced across $r = 0, 1$ (i.e., $E(y_{t_0}|r = 1) = E(y_{t_0}|r = 0)$), in which case there is no reason to use any difference-based methods.

If $\alpha_{t_1} = \alpha_{t_0} \equiv \alpha$ in GD, then $y_{it} = \alpha + \beta_t r_i + v_{it}$ and ID_{GD} is

$$\alpha + \beta_{t_1} - \alpha - \beta_{t_0} = \gamma(\alpha - \alpha) = 0 \implies \beta_{t_1} = \beta_{t_0} \ \forall\gamma\text{: } \gamma \text{ is not identified.}$$

In this case, the additive GD is more suitable.

A.2 Proof for the Figure.

If γ continue to hold for t_2, then the T group response at t_2 is

$$\begin{aligned} E + \gamma(F - C) &= E + (\frac{E - D}{C - A} + 1)(F - C) = F + (E - C) + \frac{E - D}{C - A}(F - C) \\ &= F + (E - C) + E - D \text{ as } C - A = F - C. \end{aligned}$$

If η continues to hold for t_2, then the T group response at t_2 is

$$\begin{aligned} & F + \eta(E - C) = F + (\frac{E - D}{B - A} + 1)(E - C) = F + (E - C) + \frac{E - D}{B - A}(E - C) \\ = \ & F + (E - C) + \frac{E - D}{B - A}(E - D + D - C) = F + (E - C) + \frac{E - D}{B - A}(E - D + B - A) \\ = \ & F + (E - C) + (E - D) + \frac{(E - D)^2}{B - A}. \end{aligned}$$

Since the last term is negative, $F + \eta(E - C)$ falls lower than $E + \gamma(F - C)$.

A.3 Equivalence of DG and GD

Equating DG_η to GD_γ, we get

$$E(y_{t_1} - \eta y_{t_0}|r = 1) - E(y_{t_1} - \eta y_{t_0}|r = 0) = E(y_{t_1} - y_{t_0}|r = 1) - \gamma E(y_{t_1} - y_{t_0}|r = 0).$$

Unless $E(y_{t_1} - y_{t_0}|r = 0) = 0$ in which case it is better to use the additive GD, solve this display for γ to get

$$\begin{aligned}
\gamma &= \frac{E(y_{t_1} - y_{t_0}|r = 1) - E(y_{t_1} - \eta y_{t_0}|r = 1) + E(y_{t_1} - \eta y_{t_0}|r = 0)}{E(y_{t_1} - y_{t_0}|r = 0)} \\
&= \frac{(\eta - 1)E(y_{t_0}|r = 1) + E(y_{t_1} - \eta y_{t_0}|r = 0)}{E(y_{t_1} - y_{t_0}|r = 0)}, \quad \text{cancelling } E(y_{t_1}|r = 1) \\
&= \frac{(\eta - 1)\{E(y_{t_0}|r = 1) - E(y_{t_0}|r = 0)\} + E(y_{t_1} - y_{t_0}|r = 0)}{E(y_{t_1} - y_{t_0}|r = 0)}, \quad \text{collecting terms with } \eta - 1 \\
&= \frac{(\eta - 1)\{E(y_{t_0}|r = 1) - E(y_{t_0}|r = 0)\}}{E(y_{t_1} - y_{t_0}|r = 0)} + 1
\end{aligned}$$

Hence, when $E(y_{t_1} - y_{t_0}|r = 0) \neq 0$, unless $E(y_{t_0}|r = 1) = E(y_{t_0}|r = 0)$ which does not hold in general for observational data, there exists an one-to-one relationship between η and γ; if $\eta = 1$, however, then $\gamma = 1$ always, regardless of $E(y_{t_0}|r = 1) = E(y_{t_0}|r = 0)$.

Although DG_η is equivalent to GD_γ for the two period case with $t = t_0$ and t_1, in general, the equivalence does not hold for more than two periods. This was shown already using the figure when the additional third period t_2 is the only post-treatment period. Now consider 3 periods t_0, t_1, t_2 with t_1 and t_2 being the post-treatment periods. We can obtain DG_η for (t_0, t_1) and (t_0, t_2) with a fixed $\eta \neq 1$. For this constant η, using the above display, we can obtain two different γ values for the pairs (t_0, t_1) and (t_0, t_2). This is not the same as obtaining GD_γ for the pairs (t_0, t_1) and (t_0, t_2) with the same γ.

A.4 Additive GD Version

With the additive GD, some discussions on GD become somewhat simpler. First, for Subsection 3.2 in relation to M_o $y_{it} = \alpha_t + \beta_t r_i + v_{it}$, observe

$$\gamma' = (\text{time effect at region 1}) - (\text{time effect at region 0}) = \beta_{t_1} - \beta_{t_0}.$$

The slope $\beta_{t_0} + (\beta_{t_1} - \beta_{t_0})\tau_t$ of r_i can be written as $\beta_{t_0} + \gamma'\tau_t$ to yield

$$y_{it} = \alpha_t + (\beta_{t_0} + \gamma'\tau_t)r_i + v_{it} = \alpha_t + \beta_{t_0} r_i + \gamma'\tau_t r_i + v_{it}$$

which allows for confounding interactions with $\gamma' \neq 0$. Differently from the multiplicative GD, there is no complication due to $\alpha_{t_1} = \alpha_{t_0}$

Second, for the equivalence of DG and GD in A.3, equating DG_η to $GD_{\gamma'}$ with no x, we get

$$E(y_{t_1} - \eta y_{t_0}|r = 1) - E(y_{t_1} - \eta y_{t_0}|r = 0) = E(y_{t_1} - y_{t_0}|r = 1) - E(y_{t_1} - y_{t_0}|r = 0) - \gamma'.$$

Solving this for γ' yields

$$\begin{aligned}\gamma' &= E(y_{t_1} - y_{t_0}|r=1) - E(y_{t_1} - y_{t_0}|r=0) - E(y_{t_1} - \eta y_{t_0}|r=1) + E(y_{t_1} - \eta y_{t_0}|r=0) \\ &= -E(y_{t_0}|r=1) + E(y_{t_0}|r=0) + E(\eta y_{t_0}|r=1) - E(\eta y_{t_0}|r=0) \\ &= (\eta - 1) \cdot \{E(y_{t_0}|r=1) - E(y_{t_0}|r=0)\}.\end{aligned}$$

Hence, unless $E(y_{t_0}|r=1) = E(y_{t_0}|r=0)$ which does not hold in general for observational data, there is an one-to-one relationship between γ' and η.

A.5 Dependence of Panel-Data DG Variance on η

To simplify exposition, ignore x_t and the matching procedure. Recall M_o with $v_{it} = \delta_i + u_{it}$ and its dynamic version in M_d with $\beta_y = \eta$. Assume $E(\delta_i u_{it}) = 0$ for all t, and assume that u_{it_0} and u_{it_1} are iid. Observe now

$$\begin{aligned}& y_{it} = \alpha_t + \beta_t r_i + \delta_i + u_{it} = \alpha_t + \beta_{t_0} r_i + (\eta - 1)\beta_{t_0}\tau_t r_i + \delta_i + u_{it} \\ \Longrightarrow \quad & y_{it_1} - \eta y_{it_0} = (\alpha_{t_1} - \eta\alpha_{t_0}) + (1-\eta)\delta_i + u_{it_1} - \eta u_{it_0}.\end{aligned}$$

The DG estimator $T_N(\eta)$ is

$$\begin{aligned}T_N(\eta) &\equiv \frac{1}{\sum_i r_i}\sum_i (y_{it_1} - \eta y_{it_0})r_i - \frac{1}{N - \sum_i r_i}\sum_i (y_{it_1} - \eta y_{it_0})(1 - r_i) \\ &= \frac{N}{\sum_i r_i}\frac{1}{N}\sum_i (y_{it_1} - \eta y_{it_0})r_i - \frac{N}{N - \sum_i r_i}\frac{1}{N}\sum_i (y_{it_1} - \eta y_{it_0})(1 - r_i) \\ &= \frac{1}{N}\sum_i \{\frac{N}{\sum_i r_i}(y_{it_1} - \eta y_{it_0})r_i - \frac{N}{N - \sum_i r_i}(y_{it_1} - \eta y_{it_0})(1 - r_i)\}.\end{aligned}$$

Assuming $N^{-1}\sum_i r_i \to^p \omega \equiv P(r=1)$ as $N \to \infty$, this display gives the asymptotic distribution of $T_N(\eta)$ under the H_0 of no effect $y_{it}^1 = y_{it}^0$ when the true value of η is known:

$$\begin{aligned}\sqrt{N}T_N(\eta) &\rightsquigarrow N(0, V), \quad \text{where } V \equiv \frac{E\{(y_{t_1} - \eta y_{t_0})^2 r\}}{\omega^2} + \frac{E\{(y_{t_1} - \eta y_{t_0})^2(1-r)\}}{(1-\omega)^2} \\ &= \frac{V(y_{t_1} - \eta y_{t_0}|r=1)}{\omega} + \frac{V(y_{t_1} - \eta y_{t_0}|r=0)}{1-\omega} = \frac{V(y_{t_1} - \eta y_{t_0})}{\omega(1-\omega)},\end{aligned}$$

as $y_{t_1} - \eta y_{t_0}$ is independent of r.

Examine $V(y_{t_1} - \eta y_{t_0})$:

$$\begin{aligned}& V(y_{t_1} - \eta y_{t_0}) = E[\ \{(1-\eta)\delta_i + u_{it_1} - \eta u_{it_0}\}^2\ ] = (1-\eta)^2 E(\delta^2) + E\{(u_{t_1} - \eta u_{t_0})^2\} \\ = \ & (1-\eta)^2\sigma_\delta^2 + (1+\eta^2)\sigma_u^2, \quad \text{where } \sigma_\delta^2 \equiv V(\delta) \text{ and } \sigma_u^2 \equiv V(u_t), \\ = \ & (1+\eta^2)(\sigma_\delta^2 + \sigma_u^2) - 2\eta\sigma_\delta^2 \quad (= 2\sigma_u^2 \text{ when } \eta = 1).\end{aligned}$$

Call σ_δ^2 the "between-group variance" and σ_u^2 the "within-group variance". Differentiating $V(y_{t_1} - \eta y_{t_0})$ with respect to η, we get the first derivative

$$2\eta(\sigma_\delta^2 + \sigma_u^2) - 2\sigma_\delta^2 \quad (= 2\sigma_u^2 \text{ when } \eta = 1).$$

Setting this at zero, we can see that $V(y_{t_1} - \eta y_{t_0})$ is a U-shaped quadratic function of η with its minimum at $\eta = \sigma_\delta^2/(\sigma_\delta^2 + \alpha_u^2) \leq 1$. The asymptotic-variance-minimizing η is zero when $\sigma_\delta^2 = 0$ and one when $\sigma_u^2 = 0$, which makes sense in view of the error term $(1-\eta)\delta_i + u_{it_1} - \eta u_{it_0}$. This is a 'global' finding for η. For η locally around $\eta = 1$, the first derivative is positive unless $\sigma_u^2 = 0$. That is, unless $\sigma_u^2 = 0$, DG with $\eta > 1$ tends to have a larger variance than DD. But this is not necessarily a disadvantage of DG, because the larger variance is part of the usual bias-variance trade-off; the bias is shown in the following.

The above asymptotic distribution finding is for when the true η is known. Suppose now that the wrong value of η is used, which is relevant to the sensitivity analysis as different numbers are plugged into η. Recall M_d with x_t removed and u_{it} replaced by $\delta_i + u_{it}$:

$$y_{it_1} - \beta_y y_{it_0} = (\alpha_{t_1} - \beta_y \alpha_{t_0}) + (\beta_{t_1} - \beta_y \beta_{t_0}) r_i + (1 - \beta_y)\delta_i + u_{it_1} - \beta_y u_{it_0}.$$

The intercept $\alpha_{t_1} - \beta_y \alpha_{t_0}$ gets cancelled out by the two quasi-differences in DG, but the term $(\beta_{t_1} - \beta_y \beta_{t_0}) r_i$ survives to result in

$$\sqrt{N}\{T_N(\beta_y) - (\beta_{t_1} - \beta_y \beta_{t_0})\} \rightsquigarrow N(0, V).$$

There is no change in the asymptotic variance.

REFERENCES

Altonji, J.G., T.E. Elder, and C.R. Taber, 2005, Selection on observed and unobserved variables: assessing the effectiveness of Catholic schools, Journal of Political Economy 113, 151-184.

Angrist, J. and A. Krueger, 1999, Empirical strategies in labor economics, in the Handbook of Labor Economics, Vol.3A, edited by O. Ashenfelter and D. Card, New York, Elsevier.

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, 2004, Unnatural experiments? Estimating the incidence of endogenous policies, Economic Journal 110, F672-F694.

Card, D. and D. Sullivan, 1988, Measuring the effect of subsidized training programs on movements in and out of employment, Econometrica 56, 497-530.

Cullen, J.B., B.A. Jacob, and S. Levitt, 2003, The effect of school choice on student outcomes: evidence from randomized lotteries, NBER Working paper 10113.

Heckman, J.J., R. LaLonde, and J. Smith, 1999, The economics and econometrics of active labor market programs, in Handbook of Labor Economics III, edited by O. Ashenfelter and D. Card, North-Holland.

Imbens, G.W., 2003, Sensitivity to exogeneity assumptions in program evaluation, American Economic Review (Papers and Proceedings) 93, 126-132.

LaLonde, R.J., 1986, Evaluating the econometric evaluations of training programs with experimental data, American Economic Review 76, 604-620.

Lee, M.J., 2004, Selection correction and sensitivity analysis for ordered treatment effect on count response, Journal of Applied Econometrics 19, 323-337.

Lee, M.J., 2005, Micro-econometrics for policy, program, and treatment effects, Oxford University Press.

Lee, M.J., U. Häkkinen, and G. Rosenqvist, 2007, Finding the best treatment under heavy censoring and hidden bias, Journal of the Royal Statistical Society (Series A) 170, 133-147.

Lee, M.J. and C.H. Kang, 2006, Identification for difference in differences with cross-section and panel data, Economics Letters 92, 270-276.

McKelvey, R. and W. Zavoina, 1975, A statistical model for the analysis of ordinal level dependent variables, Journal of Mathematical Sociology 4, 103-120.

Krueger, A.B. and P. Zhu, 2004, Another look at the New York city school voucher experiment, The American Behavioral Scientist 47, 658-698.

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

Neal, D., 1998, What have we learned about the benefits of private schooling? FRBNY Economic Policy Review (March), 79-86.

Neal, D., 2002, How vouchers could change the market for education, Journal of Economic Perspectives 16, 25-44.

Rouse, C.E. 1998, Private school vouchers and student achievement: an evaluation of the Milwaukee parental choice program, Quarterly Journal of Economics 113, 553-602.

Rosenbaum, P.R., 2002, Observational studies, 2nd ed., Springer-Verlag.

Rosenbaum, P.R. and D.B. Rubin, 1983, The central role of the propensity score in observational studies for causal effects, Biometrika 70, 41-55.

Shadish, W.R., T.D. Cook, and D.T. Campbell, 2002, Experimental and quasi-experimental designs for generalized causal inference, Houghton Mifflin Company.

Veall, M.R. and K.F. Zimmermann, 1996, Pseudo R^2 measures for some common limited dependent variable models, Journal of Economic Surveys 10, 241-259.