Canadian Labour Market and Skills Researcher Network

Working Paper No. 35

Estimating Treatment Effects from Contaminated Multi-Period Education Experiments: The Dynamic Impacts of Class Size Reductions

Weili Deng
Queen's University

Steven F. Lehrer
Queen's University and NBER

July 2009

CLSRN is supported by Human Resources and Skills Development Canada (HRSDC) and the Social Sciences and Humanities Research Council of Canada (SSHRC). All opinions are those of the authors and do not reflect the views of HRSDC or the SSHRC.

Estimating Treatment Effects from Contaminated Multi-Period Education Experiments: The Dynamic Impacts of Class Size Reductions

Weili Ding
Queen's University

Steven F. Lehrer*
Queen's University and NBER

Keywords: Dynamic treatment effects, contaminated experiments, class size, education production, attrition, non-compliance

JEL Code: I21 and C31.

* We wish to thank seminar participants at Harvard University, McGill University, NBER Economics of Education Fall 2003 meetings, New York University, Carnegie-Mellon University, Penn State University, Queen's University, Simon Fraser University, University of Calgary, University of California - Riverside, University of Florida, Université Laval, Economics and Education Development Conference at the Federal Reserve Bank of Cleveland, ZEW 3rd Conference on Policy Evaluation, 2008 SOLE meetings, 2006 Tar-

get Conference, 2005 CEA meetings and the 2004 NASM of the Econometric Society for comments and suggestions. We are grateful to Petra Todd for helpful discussions and encouragement at the initial stages of this project. Two anonymous referees and a coeditor made suggestions that led to a substantial improvement in the presentation of the paper. We would also like to thank Alan Krueger for generously providing a subset of the data used in the study. Lehrer wishes to thank SSHRC for research support.

Abstract

This paper introduces an empirical strategy to estimate dynamic treatment effects in randomized trials that provide treatment in multiple stages and in which various noncompliance problems arise such as attrition and selective transitions between treatment and control groups. Our approach is applied to the highly influential four year randomized class size study, Project STAR. We find benefits from attending small class in all cognitive subject areas in kindergarten and the first grade. We do not find any statistically significant dynamic benefits from continuous treatment versus never attending small classes following grade one. Finally, statistical tests support accounting for both selective attrition and noncompliance with treatment assignment.

Executive Summary

Over the past decade, many provinces have spent millions of dollars on class size reduction initiatives in the early primary grades. Proponents of these initiatives regularly draw on a subset of the published findings from Tennessee's STAR Project, a randomized intervention conducted in the late 1980s. The STAR project was conducted for a cohort of students in 79 schools over a four-year period from kindergarten through grade 3. Within each participating school, incoming kindergarten students were randomly assigned to one of the three intervention groups: small class (13 to 17 students per teacher), regular class (22 to 25 students per teacher), and regular-with-aide class (22 to 25 students with a fulltime teachers aide). However, violations to the experimental protocol were prevalent. By grade three over 50% of the subjects who participated in kindergarten left the STAR sample and approximately 10% of the remaining subjects switch class type annually.

Our empirical strategy in this paper is to reanalyze data from the STAR Project to fill the empirical gaps left behind from the numerous quantitative problems in the project's original dataset. To date, researchers have not examined the STAR Project's data as a multi period trial or accounted for the multiple violations to the experimental protocol. To accomplish these goals, this paper introduces an empirical strategy to estimate treatment effects in randomized trials that provide a sequence of interventions contaminated by various forms of noncompliance including non-ignorable attrition and selective switching between treatment and control groups at different stages of the trial. Our empirical strategy for policy evaluation of contaminated multi-period experiments creates a direct link between the structural parameters of an underlying economic model of education production, to the dynamic treatment effect estimates. This strategy can be applied to analyzing data from multi-period experiments in clinical medicine and the social sciences.

We find benefits from small class attendance initially in all cognitive subject areas in kindergarten and grade one. Yet by grade one there does not exist additional statistically significant benefits from attending small classes in both years versus attendance in one of the years. There are no statistically significant dynamic benefits from continuous treatment versus never attending small classes following grade one. Statistical tests support accounting for both selective attrition and noncompliance with treatment assignment. We investigate several potential explanations for the diminishing benefits from small class attendance in higher grades. The evidence is consistent with a story of teaching towards the bottom, in which teachers were able to identify students in the bottom of the math scores distribution and boosted their performance relative to their classmates. The evidence also suggests a trade-off between variation in academic background and class size. Taken together, the results suggest that small classes do not work unconditionally and education policymakers should exhibit caution in implementing large scale class size reductions between kindergarten and grade three.

1 Introduction

Many consider randomized experiments to be the gold standard of evaluation research due mainly to the robustness of estimators to tangential assumptions. By randomly assigning individuals to treatment, researchers can conduct an evaluation of the program that compares counterfactual outcomes without imposing strong auxiliary assumptions. However in practice, researchers regularly confront violations to the randomization protocol, complicating traditional theories of inference that require adherence to the random treatment assignment. Experiments that suffer from noncompliance with treatment assignment generate a contaminated sample in the terminology of Horowitz and Manski (1995) and defining and estimating treatment effects becomes even more challenging if there is missing outcome and background data.[1]

Numerous randomized trials in clinical medicine and the social sciences involve multiple stages of treatment receipt, during which implementation problems could proliferate as subjects may exit at the study at different periods or switch back and forth in between the treatment and control groups across time.[2] Multi-period randomized trials have the potential to address additional policy-relevant questions that extend beyond simply whether the intervention was successful as a whole. For instance one could determine when the treatment had the largest impact? How does the estimated impact of the intervention vary based on the timing of dosage? In how many periods were the treatment(s) effective?

This paper introduces an empirical strategy to estimate treatment effects in random-

ized trials that provide a sequence of interventions and suffer from various forms of noncompliance including nonignorable attrition and selective switching in between treatment and control groups at different stages of the trial. In experiments that provide a single dose of treatment, when confronted with treatment assignment noncompliance, researchers often report either an estimate of the intent to treat (ITT) parameter that compares outcomes based on being assigned to, rather than actual receipt of treatment or undertake an instrumental variables strategy. The IV estimation that uses the randomized treatment assignment as an instrumental variable for actual treatment receipt and the resulting estimate is usually interpreted as a local average treatment effect (LATE).[3] However, Frangakis and Rubin (1999) demonstrate that if the randomized intervention suffers from selective attrition, where subjects leave the study in a non-random manner, the traditional ITT estimator is biased and the IV estimator is distorted from a causal interpretation even with the assistance of a randomized instrument.

Our empirical strategy for policy evaluation of contaminated multi-period experiments permits a direct link between the structural parameters from an underlying economic model of education production to dynamic treatment effect estimates. We estimate education production functions using a sequential difference in difference estimator to control for selective switching and account for non-ignorable attrition using inverse probability weighting. That is, we map a set of structural parameters obtained from estimating one of the most commonly used model of human capital accumulation into a statistical estimator

that has a causal interpretation. This empirical strategy could also be readily applied to estimate the full sequence of dynamic treatment effects from interventions where the outcome is posited to be generated from a cumulative process such as health human capital or asset accumulation over the lifecycle.

We use data from Tennessee's highly influential class size experiment, Project STAR to illustrate our empirical strategy. This experiment was conducted for a cohort of students in 79 schools over a four-year period from kindergarten through grade 3. Within each participating school, incoming kindergarten students were randomly assigned to one of the three intervention groups: small class (13 to 17 students per teacher), regular class (22 to 25 students per teacher), and regular-with-aide class (22 to 25 students with a full-time teacher's aide). However, violations to the experimental protocol were prevalent. By grade three over 50% of the subjects who participated in kindergarten left the STAR sample and approximately 10% of the remaining subjects switch class type annually. To the best of our knowledge, an examination of the data as the result of a sequence of contaminated treatment interventions has not been explored.[4]

This paper is organized as follows. In Section 2, we describe the causal parameters of interest in multi-period experiments and introduce an empirical framework that builds on the standard economic model of human capital accumulation. The assumptions underlying our identification strategy are discussed and the estimation approached is detailed in this section. We demonstrate that both our structural parameter and treatment effect

parameters are nonparametrically identified. Section 3 presents a description of the data used in our analysis. Our results are presented and discussed in Section 4. We find benefits from small class attendance initially in all cognitive subject areas in kindergarten and grade one. Yet by grade one there does not exist additional statistically significant benefits from attending small classes in both years versus attendance in one of the years. There are no statistically significant dynamic benefits from continuous treatment versus never attending small classes following grade one. A concluding section summarizes our findings and discusses directions for future research.

2.0 Causal Parameters of Interest

In the context of the STAR class size experiment, we refer to being in small classes as receiving treatment, attending either regular or regular with aide classes as being in the control group.[5] We use $S_t = 1$ to denote attending a small class in grade t and $S_t = 0$ as being in a regular class. At the completion of each grade t, a student takes exams and scores A_t (potential outcomes: A_{1t} if attending a small class and A_{0t} if attending a regular class). An evaluation problem arises since we cannot observe A_{1t} and A_{0t} for the same individual.

Project STAR was conducted to evaluate the effect of class size on student achievement to determine whether small class size should be extended to the schooling population as a whole. In a single period experiment, the relevant parameter of policy interest is the average treatment effect (ATE) $\triangle_{ATE_t} = E(A_{1t} - A_{0t})$ or in its conditional form $E(A_{1t} -$

$A_{0t}|X)$ where X are characteristics that affect achievement. However, due to the non-mandatory compliance nature of the Project STAR experiment, each year the actual class type a student attends may differ from their initial assignment.[6] When individuals self-select outside of their assigned groups, risks rise that the groups may no longer be equivalent prior to treatment and the experimental approach is not able to identify the ATE,[7] in which case researchers either report an ITT or conduct an IV analysis.[8]

Project STAR was carried out on a cohort of students beginning in kindergarten through the third grade. The standard evaluation problem becomes more challenging with multiple stages of treatment receipt as the number of potential outcomes increases. For instance, with two stages of treatment, an individual could complete one of four possible sequences $[(S_{i2}=1, S_{i1}=1), (S_{i2}=1, S_{i1}=0), (S_{i2}=0, S_{i1}=1), (S_{i2}=0, S_{i1}=0)]$. An individual's outcome at the conclusion of the second period can be expressed as

$$A_{i2} = S_{i1}S_{i2}A_i^{11} + (1-S_{i1})S_{i2}A_i^{01} + S_{i1}(1-S_{i2})A_i^{10} + (1-S_{i1})(1-S_{i2})A_i^{00} \qquad (1)$$

where A_i^{11} indicates participation in small classes in both periods, A_i^{10} indicates small class participation only in the first period, etc. It is clear that an individual who participated in both periods (A_i^{11}) has three potential counterfactual sequences to estimate (A_i^{01}, A_i^{10} *and* A_i^{00}) if the four paths are all the sequences an individual can take. In our multi-period intervention framework answers to many hotly debated questions, such as when class size reductions are most effective or whether small class treatment in early grades provide any additional benefits in later grades can be obtained.

In a multi-period setting, the relevant causal parameters of policy interest are the full sequence of dynamic average treatment on the treated parameters. Following Lechner (2004), we formally define $\tau^{(x,y)(v,w)}(x,y)$ the dynamic average treatment effect for the treated parameter. $\tau^{(x,y)(v,w)}(x,y)$ measures the average difference in outcomes between their actual sequence (x,y) with potential sequence (v,w), for individuals who participated in program x in period 1 and program y in period 2. For example, $\tau^{(1,1)(0,0)}(1,1)$ is an estimate of the average cumulative dynamic treatment effect for individuals who received treatment in both periods. Similarly, $\tau^{(1,1)(1,0)}(1,1)$ is an estimate of the effect of receiving treatment in the second year for individuals who received treatment in both periods, and $\tau^{(0,1)(0,0)}(0,1)$ is the effect of receiving treatment in the second period for individuals who received treatment only in period two.

2.1 Empirical Model

We construct dynamic treatment effect for treated parameters (DTET) from estimates of the structural parameters of an education production function. Following Ben-Porath (1967) and Boardman and Murnane (1979), we view the production of education outcomes as a cumulative process that depends upon the potential interactions between the full history of individual, family and school inputs (captured in a vector X_{ijt} in year t), class size treatments, innate abilities and independent random shocks $(\epsilon_{iT}...\epsilon_{i0})$. Formally, child i in school j gains knowledge as measured by a test score at period T :

$$A_{ijT} = h_T(X_{iT}...X_{i0}, S_{jT_T}...S_{jT_o}, v_i, \epsilon_{iT}...\epsilon_{i0}) \tag{2}$$

where h_T is an unknown twice differentiable function. Note v_i is included to capture unobserved time invariant individual attributes.

In our empirical analysis, we first linearize the production function at each time period. An individual's achievement outcome in period one is expressed as

$$A_{i1} = v_i + \beta_1' X_{i1} + \beta_{S1}' S_{i1} + \varepsilon_{i1} \tag{3}$$

where v_i is a individual fixed effect. Similarly in period two achievement is given as

$$A_{i2} = v_i + \alpha_2' X_{i2} + \alpha_1' X_{i1} + \alpha_{S2}' S_{i2} + \alpha_{S1}' S_{i1} + \alpha_{S12}' S_{i2} S_{i1} + t_2 + \varepsilon_{i2} \tag{4}$$

and t_2 reflects period two common shock effects. Since nearly all of the explanatory variables in equations (3) and (4) are discrete dummy variables the only restrictive assumption by linearization is the additive separability of the error term.[9] This implementation allows the effects of observed inputs and treatment receipt on achievement levels to vary at different grade levels.[10] We also allow the effect of being in a small class in the first year (S_{i1}) on second period achievement (A_{ij2}) to interact in unknown ways with second year class assignment (S_{i2}). First differencing the achievement equations generates the following system of equations

$$\begin{aligned} A_{i2} - A_{i1} &= \alpha_2' X_{i2} + \alpha_{S2}' S_{i2} + \alpha_{S12}' S_{i2} S_{i1} + t_2 + (\alpha_1 - \beta_1)' X_{i1} + (\alpha_{S1} - \beta_{S1})' S_{i1} + \varepsilon_{i2}^* \quad (5) \\ A_{i1} &= \beta_1' X_{i1} + \beta_{S1}' S_{i1} + \varepsilon_{i1}^* \end{aligned}$$

where $\varepsilon_{i2}^* = \varepsilon_{i2} - \varepsilon_{i1}$ and $\varepsilon_{i1}^* = \upsilon_i + \varepsilon_{i1}$. As this is a triangular system of equations, full information maximum likelihood parameter estimates are equivalent to equation by

equation OLS which does not impose any assumptions on the distribution of the residuals. Consistent and unbiased structural estimates of β_{S1} and of the teacher characteristics in the X_{i1} matrix can be obtained since subjects and teachers were both randomized between class types in kindergarten and compliance issues did not arise until the following year.[11]

To estimate the DTET defined in the preceding subsection, our approach builds on Miquel (2003), who demonstrates that such a conditional difference-in-differences approach of the achievement equations can nonparametrically identify the causal effects of sequences of interventions. The structural parameter estimates from equation (5) are used to calculate the full sequence of dynamic effects as follows:

$$\begin{aligned} \tau^{(1,1)(0,0)}(1,1) &= \alpha_{S1} + \alpha_{S2} + \alpha_{S12} \\ \tau^{(1,1)(1,0)}(1,1) &= \alpha_{S2} + \alpha_{S12} \\ \tau^{(0,1)(0,0)}(0,1) &= \alpha_{S2} \end{aligned} \tag{6}$$

Dynamic variants of the straightforward assumptions of common trend, no pretreatment effects and a common support condition are required to obtain causal parameters.[12] It is straightforward to extend this strategy to T periods.

While concerns regarding non-compliance with treatment assignment are addressed by controlling for the history of observed inputs and assuming the effects of individual unobserved heterogeneities which include factors such as parental concern over their child's development are fixed over short time periods, attrition remains a concern. Define $L_{t+1} = 1$ to indicate that a subject leaves a STAR school and attends a school elsewhere after

completing grade t, if she remains in the sample next period $L_{t+1} = 0$.[13] Attrition may be due to exogenous and endogenous observables that are observed prior to attrition. If only selective attrition based on observables is present, the attrition probability is independent of the dependent variable (and hence unobserved factor), which implies that $Pr(L_t = 0|A_t, X_t) = Pr(L_t = 0|X_t)$. As such, estimates can be re-weighted and the conditional population density $f(A_t|X_t)$ can be inferred from $g(A_t|X_t, L_t = 0)$ even though A_t is observed only if $L_t = 0$.

Consider the probability of attrition model implies a process of the form

$$L^*_{t+1} = 1\{\alpha' Z_{it} + w_{it} \geq 0) \tag{7}$$

where L^*_{t+1} is a latent index and $L_{t+1} = 1$ if $L^*_{t+1} \geq 0$, w is a mean zero random variable whose c.d.f. is F_w, t is the period being studied and Z_{it} is a matrix of predetermined variables (A_{it}, S_{it}, X_{it}) that are observed conditional on $L_t = 0$ and also include lagged dependent variables (A_{t-s}) as well as past test scores in all other subject areas.[14] The probability of staying in the sample $Pr(L_{it+1} = 0|A_{it}, S_{it}, X_{it}) = F_w(-\alpha' Z_{it})$, and in our analysis we begin by assuming that w_{it} follows a symmetric distribution to estimate the probabilities of remaining in the experiment $\hat{p}_{it}$. By reweighting observations using $\hat{p}_{it}$when estimating equation (5), reexpresses the system of equations as

$$\begin{aligned} \frac{A_{i2} - A_{i1}}{\hat{p}_{i1}} &= \frac{\alpha_2' X_{i2} + \alpha_{S2}' S_{i2} + \alpha_{S12}' S_{i2} S_{i1} + t_2 + (\alpha_1 - \beta_1)' X_{i1} + (\alpha_{S1} - \beta_{S1})' S_{i1} + \varepsilon^*_{i2}}{\hat{p}_{i1}} \\ A_{i1} &= \beta_1' X_{i1} + \beta_{S1}' S_{i1} + \varepsilon_{i1} \end{aligned} \tag{8}$$

which generates $\sqrt{N}$ consistent estimates that are asymptotically normal.[15] Correcting for selection on observables in the panel by inverse probability weighting reduces the amount of residual variation in the data due to attrition. Since attrition in the STAR sample is an absorbing state, the weights used in estimation of equation (8) for grades two and three ($\hat{r}_i^2 = \hat{p}_{i2} * \hat{p}_{i1}$and $\hat{r}_i^3 = \hat{p}_{i3} * \hat{p}_{i2} * \hat{p}_{i1}$) are simply the product of all current and past estimated probabilities, where $\hat{p}_{is}$ are estimated probabilities for staying in the sample for period s from a logit regression using all subjects in the sample at $s-1$. We can include school effects to the estimating equations, however, identification of school effects will only come from the limited number of school switchers.

3.0 Project STAR Data

Project STAR was a large scale experiment that initially randomized over 7,000 students in 79 schools into one of the three intervention groups: small class (13 to 17 students per teacher), regular class (22 to 25 students per teacher), and regular-with-aide class (22 to 25 students with a full-time teacher's aide) as the students entered kindergarten.[16] Teachers were also randomly assigned to the classes they would teach.[17] The experiment continued until the students were in grade three and academic performance measures were collected at the end of each year. In our analysis, we use scaled scores from the Reading, Mathematics and Word Recognition sections of the Stanford Achievement Test since that scoring system allows us to use differences in scaled scores as measures to track development between grades. We investigate the impact of small classes on each outcome

separately since one may postulate that the treatment could be more effective in subject areas such as mathematics where classroom instruction is used as opposed to group instruction for reading.

In our empirical analysis, we include only the sample of students who participated in the STAR experiment starting in kindergarten. Fewer than half of the kindergarten students participated in all four years of the experiment (3085 out of 6325 students).[18] Each year there were also movements between small and regular classes for this cohort of students. Figure 1 presents the number of students on each potential treatment path at each graded level. Excluding attrition there is support for all eight sequences in grade two and fourteen of the sixteen possible sequences in grade three. The large number of transitions illustrated in Figure 1 motivate our empirical strategy developed in the preceding section.

3.1 Sample Construction and Selective Attrition

We did not pool the kindergarten sample with the refreshment samples (students who joined the experiment after kindergarten) since we find evidence from regressions that i) students did not leave the Project STAR experiment in a random manner, and ii) subsequent incoming groups were not conditionally randomly assigned within each school. Specifically, to examine conditional random assignment of the refreshment sample for each group of students entering the experiment after kindergarten we conducted straightforward regressions of a random treatment assignment indicator (M_{ijT}) on individual

characteristics and school indicators as follows

$$M_{ijT} = \gamma' X_{ijT} + \upsilon_j + e_{ijT} \tag{9}$$

where $M_{ijT} = 1$ if a student is initially assigned to a small class when she enters a school in the STAR sample and $M_{ijT} = 0$ otherwise. If students are assigned randomly there should be no evidence of a systematic differences in baseline characteristics (as well as unknown confounders) between the treatment and control group.

Estimates of equation (9) using the sample of only incoming students in each grade are presented in the top panel of Table 1. The results demonstrate that incoming students to the experiment that were on free lunch status were more likely to be assigned to the control group in both grades one and three. Coupled with the movements of the existing students in the sample there were significant different in student characteristics between small and regular classes. Estimates of equation (9) using the full sample of students in each grade are presented in the bottom panel of Table 1. Students who are white or Asian, female and not on free lunch status are statistically more likely to be currently attending a small class in each year following Kindergarten.

To examine whether the subjects left the experiment in a non-random manner, we first test for attrition due to observables using the procedure developed in Becketti, Gould, Lillard and Welch (1988). We estimate the following equation

$$A_{ij1} = \beta' X_{ij1} + \beta'_L L_{ij} X_{ij1} + \upsilon_j + \varepsilon_{ij1} \tag{10}$$

where A_{ij1} is the level of educational achievement for student i in school j in the first year, X_{ij1} is a vector of initial school, individual and family characteristics, L_{ij} is an indicator for *subsequent* attrition $L_{ij} = 1$ if $L_{ijs} = 1$ for any $s = 2...T$), υ_j is included to capture unobserved school specific attributes and ϵ_{ij1} captures unobserved individual factors. Selection on observables is non-ignorable if β_L is statistically significant, indicating that individuals who subsequently leave the STAR experiment were systematically different from those who remain in terms of initial behavioral relationships.[19]

Table 2 presents estimates of equation (10) and Wald tests presented in the third row from the bottom of Table 2 indicate that the β_L coefficient vector is significantly different for attritors from non-attritors in all subject areas. Further, the second row from the bottom of Table 2 demonstrates that the joint effect of attrition on all student characteristics and class type is significantly different from zero in all three subject areas. Examining the individual coefficient estimates in Table 2 notice that the attrition indicator is significantly negatively related to test scores in all three subject areas indicating that subsequent attritors scored significantly lower on average in all kindergarten cognitive examinations. Students on free lunch status that left scored significantly lower than free lunch students who remained in the sample in mathematics. Interestingly, female attritors out performed female non-attritors in kindergarten in all subject areas but the magnitude is small. In both mathematics and word recognition attritors received half of the average gains of being in a small class. Since non-attritors in small classes, obtained larger gains in

kindergarten, future estimates of the class size effect may be biased upwards if attrition is not controlled for. As there is no evidence that attrition patterns differed between schools in Tennessee that participated and did not participate in the STAR experiment, concerns regarding selection on unobservables are reduced.[20]

4.0 Empirical Results

Our structural estimates of the causal effects of reduced class size from estimating equation system (8) are presented in Table 3. In kindergarten and grade one small class attendance ((S_{iK}) *and* (S_{i1})) has positive and significant effects in all three subjects areas. However, there does not exist additional (nonlinear) benefits from attending small classes in both years ($S_{iK}S_{i1}$). After grade one, no significantly positive effects of small class exists ($P(t) \leq 10\%$) with the exception of grade two math. The average small class effects in grade three (S_{i3}) are significantly ($\leq 10\%$) negatively related to achievement in all three subjects.

Table 4 presents some estimates of the dynamic average treatment effect for the treated in which we compare the sequences with the largest number of observations. In grade one, the set of DTETs suggest that the largest gains in performance in all subject areas accrue for students who attended small classes either in kindergarten or in grade one ($\tau^{(0,1)(0,0)}(0,1)$ or $\tau^{(1,0)(0,0)}(1,0)$). Benefits from attending small classes in both kindergarten and grade one versus attendance in either but not for both of these years ($\tau^{(1,1)(0,1)}(1,1)$ or $\tau^{(1,1)(1,0)}(1,1)$) are statistically insignificant. While the economic signif-

icance of attending a small class in grade one alone is slightly greater in all subject areas than attendance in kindergarten alone ($\tau^{(0,1)(0,0)}(0,1) > \tau^{(1,0)(0,0)}(1,0)$), there does not exist a statistically significant difference between either sequence ($\tau^{(0,1)(1,0)}(0,1)$). From a policy perspective the results do not lend support for providing small class as continuing treatments.

The pattern in higher grades presents several additional insights into the effectiveness of reduced class size. The dynamic benefits from continuous treatment versus never attending small classes ($\tau^{(1,1,1)(0,0,0)}(1,1,1)$ and $\tau^{(1,1,1,1)(0,0,0,0)}(1,1,1,1)$) become both statistically and economically insignificant in all subject areas. In grade one, approximately 250 students substituted into the treatment and received positive benefits. Continuing along this path and remaining in small classes in higher grades did not provide any additional benefits for those students as both $\tau^{(0,1,1)(0,0,0)}(0,1,1)$ and $\tau^{(0,1,1,1)(0,0,0,0)}(0,1,1,1)$ are statistically insignificant. Further, their economic significance is smaller than $\tau^{(0,1)(0,0)}(0,1)$. Similar to Krueger (1999) we find that students received large benefits the first year they spent in a small class in all subject areas in grade one and in math in grade two. However, we find that students who entered small classes for the first time in grade three achieved significant losses from attending a small class ($\tau^{(0,0,0,1)(0,0,0,0)}(0,0,0,1)$) in all subject areas. Finally, students who switched into small class for the first time in grade two did not have statistically significant gains on reading and word recognition ($\tau^{(0,0,1)(0,0,0)}(0,0,1)$).

This study differs from past research on Project STAR not solely through the focus

of treating the experiment as a multi-period intervention but also in accounting for both attrition due to observables and the possibility that other forms of noncompliance are due to unobservables. Tables 5 and 6 presents results from specification tests to determine if we should statistically account for non-compliance and attrition. Results from DuMouchel and Duncan (1983) tests presented in Table 5 support accounting for attrition due to observables in all subject areas and all grade levels at conventional levels ($P(F) \leq 5\%$) in reading and mathematics and below the 20% level in word recognition. Likelihood ratio tests presented in Table 6 are conducted to determine whether one should include v_i, which proxies for the possibility that noncompliance may be due to unobservables. In all subject areas and all grades the Null hypothesis is rejected, supporting the presence of individual unobserved heterogeneity. Hausman tests between estimates of the simpler system of equations that did not include v_i and equation (8) reject the restriction that $v_i = 0$, lending further support that noncompliance of treatment assignment is selective.

4.1 Discussion

The estimates in Tables 3 and Table 4 provide a richer picture of the impacts from class size reductions. A significant impact from smaller classes appears in kindergarten. Following kindergarten, the positive effects of smaller classes in grade one accrue only for those students who made a transition between class types. Students who substituted into small classes and dropped out of small classes both scored significantly lower than their grade one classmates in each kindergarten subject. Additionally these students

received a significantly larger improvement in grade one achievement compared to their grade one classmates as well as their kindergarten classmates.[21] Several of our results are consistent with Hanushek (1999) in suggesting that there was an erosion of the early gains from small class attendance in later grades. In this subsection, we investigate several possible explanations for the diminishing benefits from small classes and present evidence that the behavior and characteristics of the students who did not comply with treatment assignment are primarily responsible for the changes in the sign and significance of the DTET.

We first conduct a closer examination of the students who switched class types at the time of their initial switch. Using classroom level regressions we compared these students who either dropped out of, or substituted into, small classes with their new classmates based on prior exam performance by subject area. In grades one and two, students who joined small classes scored significantly lower upon entry than their new classmates with the exception of reading for those who substituted in grade two. In grade three, students who switched into small classes for the first time scored significantly higher on past exams than their new classmates. Thus, the academic background of these individuals who switched class type changed over time. Interestingly the subsequent achievement of these switching individuals relative to their new classmates also exhibited a statistically significant pattern whose direction changed when the relative academic backgrounds of the switching students improved over time.[22] In grades one and two, students who switched

class type achieved significantly greater growth on mathematics.[23] However, in grade three, students who switched into a small class for the first time achieved a significantly smaller gain in their math score relative to their new classmates. A potential explanation for this pattern of results is that teachers were targeting the weaker students in the class.

Coleman (1992) suggests that the focus of US education is on the bottom of the distribution and it is much easier for teachers to identify weaker students in mathematics than other subject areas. The major challenge in formally investigating this claim is separating the amount of test score gains from teachers' characteristics from a statistical tendency called "regression to the mean", which is created by non-random error in the test scores. To address this issue we classified the five students in each grade one classroom that had the lowest scores on kindergarten tests in each subject as being a "weak" student in that area. We included an indicator variable for being one of these "weak" students in the classroom in regression equations to explain growth in performance controlling for the full history of teacher, family and student characteristics. Using multiple regression we separately examined whether being a "weak" student in math or reading or word recognition led to larger gains in test performance in all subject areas. Consistent with the regression to the mean argument students who were "weak" in mathematics and word recognition received larger gains in performance relative to their classmates in these subject areas. In contrast, being a "weak" student in reading significantly reduced gains in reading performance in grade 1. Supporting Coleman's hypothesis, we found that the

"weak" students in math also achieved larger gains in their classroom in both reading and word recognition, but the same gains do not exist for "weak" students in reading or word recognition.[24] When we focus our examination on students who switched class type, we find that they only achieved benefits from switching into small classes if their past performance in math was significantly lower.

Noncompliance with treatment assignment also resulted in an increased variation of student background within classrooms in higher grades. Specifically, small classes in grades two and three have significantly more variation in incoming performance in math and reading than regular classes as many "weak" students made transitions from regular to small classes.[25] Faced with less variation in the incoming knowledge of their classmates, linear regressions demonstrate that students in regular classes were able to achieve significantly larger gains in math and reading in grade two and in math in grade three.[26] This result is not driven by the subset of students who switched class type, as both simple t-tests and multiple regression results that compare the experience of the subset of students who always complied with their assignment, (i.e. always versus never attended small classes) indicate that those students who never attended small classes experienced significantly larger growth in mathematics both in grade two and grade three. These students also had greater gains on the second grade reading exam.[27] As the heterogeneity in academic background became smaller over time in regular classes, the dynamic benefits of small class attendance vanished and even reversed in some subjects. Consistent with

this explanation, we do not find any evidence for significant differences in performance on word recognition exams, the only subject in which there is no evidence for significant differences in the variation of prior performance. Taken together, the patterns reported in Tables 3 and 4 for grades two and three might suggest a trade-off between variation in incoming student performance and class size.[28] Unfortunately, we cannot formally investigate this trade-off because the peer compositions are no longer exogenous in higher grades.[29]

The benefits occurring to students who did not comply with treatment assignment following kindergarten seems to run counter to the hypothesis that students benefit from environmental stability. We conducted an examination of the effects of environmental stability on students in small classes in grade one.[30] In each grade one small class, we first identified members of the largest subgroup of students who were taught by the same kindergarten teacher. From OLS regressions that control for the full history of teacher, family and student characteristics we found that students who were members of the largest subgroup had significantly smaller gains relative to their classmates in mathematics (coeff.=-6.129, s.e. 2.714) and word recognition (coeff.=-4.524, s.e. 3.008) but no significant differences in reading.[31] These results do not support environmental stability arguments, nor do they directly contradict the stability hypothesis since peer groups (classmates) were no longer exogenously formed after kindergarten.

To check the robustness of our estimates in Tables 3 and 4, we consider two strategies

that increase the statistical power of the structural parameter and dynamic treatment effect estimates and a strategy that relaxes implicit parametric assumptions in the attrition model.[32] Specifically we i) ignore potential nonlinear impacts of the small class treatments in equation (8),[33] ii) relax the identification assumptions for the attrition model allowing us to use a larger sample,[34] and iii) relax the parametric assumptions used to estimate equation (7).[35]

The results of these robustness check (available upon request) suggest that the differences in our findings from earlier work are unlikely due to statistical power or parametric assumptions. In higher grades, kindergarten small class attendance (S_{iK}) is positively related to performance in grade two reading and grade three reading and word recognition examinations. Whereas, attendance in small classes in grade one (S_{i1}) is either negatively related or unrelated to performance in both grades two and three. The results suggest that there could be some small positive effects from attending a small class in kindergarten in reading and word recognition in higher grades. For mathematics, the results appear to suggest that small class attendance in both kindergarten and grade two may have some lasting impacts. As before, we find that nearly every path of multiple receipts of treatment in the higher grades is not significantly related to achievement in any subject area. Overall, these results suggest that the benefits of attending a small class early on are of small magnitude and a single dose in kindergarten yields most of the benefit. The substantial heterogeneity in the treatment effects makes it important to understand

the reason why small classes work when they are effective, and similarly understand the explanations for their failures. For example, more understanding of the nature of class size and relationship with teaching practices is needed. To summarize the results suggest that small classes do not work consistently and unconditionally.

5 Conclusion

Randomized trials often suffer from a number of complications, notably noncompliance with assigned treatment and missing outcomes. These problems could potentially proliferate in longitudinal experiments that expose subjects to treatment at different points in time. In this paper we introduce an empirical strategy to estimate treatment effects in randomized trials that provide a sequence of interventions and suffer from various forms of noncompliance including nonignorable attrition and selective switching in between treatment and control groups at different stages of the trial. Our empirical strategy for policy evaluation also permits a direct link between the structural parameters from an underlying economic model of education production to dynamic treatment effect estimates.

To illustrate our empirical strategy we use data from the highly influential randomized class size study, Project STAR. We find benefits from small class attendance initially in all cognitive subject areas in kindergarten and the first grade. We do not find any statistically significant dynamic benefits from continuous treatment versus never attending small classes in either the second or third grade. Statistical tests support accounting for both selective attrition and noncompliance with treatment assignment. Finally, we investigate

several potential explanations for the diminishing benefits from small class attendance in higher grades. The evidence is consistent with a story of teaching towards the bottom, in which teachers were able to identify students in the bottom of the math scores distribution and boosted their performance relative to their classmates. The evidence also suggests a trade-off between variation in academic background and class size. Examining these explanations in greater detail present an agenda for future research.

Notes

[1]An experimental study with endogenously censored outcomes within a contaminated sample produces a corrupted sample, in the terminology of Horowitz and Manski (1995). Barnard, Du, Hill and Rubin (1998) coined the term "broken randomized experiments" to describe such studies that experience more than one partially uncontrolled factor (i.e. non-compliance and missing data) in implementation. Frangakis and Rubin (2002) developed a Bayesian approach to estimate alternative causal parameters from broken randomized experiments. Our approach differs based on statistical assumptions imposed, causal parameters estimated and has a direct link to the structural parameters from an economic model of education production.

[2]The study of causal effects from a sequence of interventions is limited even in the case of perfect compliance. Only recently in economics, Lechner and Miquel (2005), Lechner (2004) and Miquel (2002, 2003) examine the identification of dynamic treatment effects under alternative econometric approaches when attrition is ignorable. The original investigation on treatment effects explicitly in a dynamic setting can be traced to Robins (1986). More recent developments in epidemiology and biostatistics can be found in Robins et al. (2000) and Yau and Little (2001). In these papers, subjects are required to be re–randomized each period to identify the counterfactual outcomes.

[3]It obtains this causal interpretation provided a series of assumptions detailed in Imbens and Angrist (1994) as well as Angrist, Imbens and Rubin (1996) are satisfied.

[4]Most published findings from this study have reported large positive impacts of class size reduction on student achievement, a subset of which have noted and attempted to address complications due to missing data and noncompliance with the randomly assigned treatment that occurred during implementation. For example, Krueger (1999) presents IV estimates to correct for biases related to deviations from treatment assignment.

[5]Following Finn et al. (2001) and Krueger (1999) our control group consists of regular class with and without teacher aides, as these studies (among others) report that the presence of a teacher aide did not significantly impact student test scores. However, to date whether teaching aides have impacts on academic performance in regular classes has not been examined by accounting for multiple stages of treatment and estimating dynamic treatment effects.

[6]Detailed discussions of the consequences of different forms of non-compliance with treatment assignment in single period experiments can be found in Heckman Smith and Taber (1998), Heckman, Hohmann Smith and Khoo (2000) and Section 5.2 of Heckman, Lalonde and Smith (2001).

[7]Researchers (i.e. Manski (1990), Balke and Pearl (1997), among others) have demonstrated that the ATE is partially identified.

[8]Balke and Pearl (1997) demonstrate that in studies which experience noncompliance with treatment assignment, both ITT and IV point estimates are potentially misleading, as they could lie outside the theoretical bounds constructed for the ATE. Ding and Lehrer

(2008) use the same data as in this study and consider several alternative strategies that place bounds on ATE, comparing them to the ITT and IV point estimates. The construction of alternative sets of bounds relaxes alternative identifying assumptions also allows the reader to ascertain the robustness of the conclusions to the maintained assumptions.

[9]To identify the structural parameter we do not need to linearize the education production function. Assuming that the unobserved factors enter additively, and that i) the unobserved components ν_i, ε_{i1} are independent of S_{i1}, ii) $(\varepsilon_{i1}, \varepsilon_{i2})$ is independent of $(X_{i1}, S_{i1}, X_{i2}, S_{i2})$ and iii) t_2 is a constant; the structural parameter of class type are non-parametrically identified. Chesher (2003) additionally points out that a local insensitivity assumption is needed to achieve local identification of the partial derivatives of structural functions in a triangular system of equations.

[10]We place no restrictions such as forcing the depreciation rate to be constant across all inputs in the production process, which is generally done when estimating education production functions. However, we assume that the effect of unobserved inputs is constant between successive grades. The validity of this assumption was tested using a IV procedure developed in Ding and Lehrer (2004) and supported in both grades 2 and 3.

[11]The importance of randomization and the fact that compliance was near perfect in kindergarten is crucial to our identification strategy. While the possibility exists that some students were switched from their randomly assigned class to another class before kindergarten started, Krueger (1999) examined actual enrollment sheets that were compiled in

the summer prior to the start of kindergarten for 1581 students from 18 participating STAR schools and found that only one single student in this sample who was assigned a regular or regular/aide class enrolled in a small class.

[12]The common support assumption ensures that there are comparable individuals in each of the counterfactual sequence. The common trend assumption assumes that the sole difference before and after is due to treatment across groups as in the absence of treatment the comparing groups would have in expectation similar gains in academic performance. The no pre-treatment assumption requires that there is no effect of the treatment on outcomes at any point in time prior to actual participation.

[13]Fitzgerald, Gottschalk and Moffitt (1998) describe specification tests to detect attrition bias and methods to adjust estimates in its presence.

[14]Identification is obtained from historical test scores.

[15]However, the asymptotic variance is conservative since it ignores the fact that we are weighting on the estimated and not the actual $\hat{p_{i1}}$. See Wooldridge (2002) for details and a discussion of alternative estimation strategies. The full set of results is available by request where the asymptotic covariance matrix of the second step estimator is computed using the results of Newey (1984) that account the use of generated regressors.

[16]Students were assigned to a class type based on their last name using a centrally prepared algorithm and school specific starting value.

[17]A potential concern is whether the teachers in this study altered their behavior in

response to treatment assignment. It is reasonable to speculate that teachers may have selected specific instruction methods that could either reinforce or counteract the impacts of small classes. Unfortunately, data from Project STAR process evaluations remains publically unavailable to determine whether teachers selectively altered their behavior (e.g. we do not have any evidence related to John Henry or Hawthorne effects in the study). Throughout this paper, we are implicitly assuming that teachers did not have a behavioral response to treatment assignment. In Ding and Lehrer (2008), we demonstrate that the bounds for the ATE do not exhibit major changes in most grades and subject areas when we relax assumptions related to whether teachers have a behavioral response to the study, suggesting that any bias is fairly small. We are grateful to an anonymous referee for pointing out this potential limitation.

[18]For the full kindergarten sample, a linear probability model regression of subsequent attrition on initial class assignment yields a statistically significant impact of class type. The attrition rate also varied significantly by class type across schools.

[19]Fitzgerald et al. (1998) demonstrate that this test is simply the inverse of examining whether past academic performance significantly affects the probability of remaining in the study from estimating equation (7).

[20]Information on students from similar non-participating schools has been collected.

[21]These findings are obtained from within classroom regressions that control for kindergarten and grade one student, family and teacher characteristics.

[22]Since scaled scores are developmental they can be used to measure growth across grades within the same test subject area allowing us to make these comparisons.

[23]Further, these growth rates were significantly larger than those achieved by their kindergarten classmates who did not switch in grade one.

[24]Our results are robust to several alternative definitions of being a "weak" student. We also defined being a "weak" student as having the lowest or one of the three or four lowest scores in the classroom. Note, if regression to the mean were the prime explanation we should expect to see this improvement not only for students with low incoming math scores. However, the improvement in subsequent performance in all subject areas does not exist for "weak" students in the other subject areas.

[25]T-tests on the equality of variances in incoming test scores indicate significantly larger variation in small classes in mathematics in grades two and three and in grade two reading. Variation may influence student performance through teaching methods as instructors may face additional challenges engaging students at different levels.

[26]Regressions including school indicators demonstrate that performance gains in reading between grades one and two (coeff.=-2.54, s. e.=1.05) and gains in mathematics between grades one and two (coeff. =-2.22, s. e.=1.11) and between grades two and three (coeff. =-2.21, s. e.=0.88) were significantly lower in small classes.

[27]The regressions include school indicators as well as student and teacher characteristics. The effect (and standard error) of always attending a small class (relative to never) is -4.18

(1.46) in grade two reading gains and -2.75 (1.35), -2.18 (1.28) in grade two and grade three mathematics gains respectively. Note in grade one, there are positive and significant gains for always attending a small class in reading and word recognition which explains the dynamic benefits at that time. The full set of results are available from the authors.

[28]Our findings are consistent with evidence on elementary school students presented in Hoxby (2000a) and Hoxby (2000b) who exploited natural variation in age cohorts in the population and found evidence that class size does not affect student achievement in Connecticut and peer group composition affects achievement in Texas respectively.

[29]The dataset in its current form does not allow for control of the endogenous peer formation after kindergarten.

[30]We do not analyze students in regular classes since they were re-randomized within schools between classes with and without aides following kindergarten.

[31]Multiple regressions using the number of current classmates who were also taught with the same kindergarten teacher (instead of a simple indicator variable) also find significantly smaller gains in mathematics (coeff.=-1.797, s.e. 0.572) and word recognition (coeff.=-1.179, s.e. 0.572) for each additional former classmate.

[32]We also considered estimating the ITT for the subset of subjects who complied with their assignment throughout the study. This removes all selective switchers from the analysis and focuses attention on the two main pathways.

[33]This model is less flexible than that estimated in Table 3 and implicitly places several

equality restrictions on several dynamic treatment effect paths. For example, in grade two $\tau^{(0,1,1)(0,0,1)}(0,1,1) = \tau^{(0,1,0)(0,0,0)}(0,1,0)$. We constructed F tests on the joint significance of the non-linear interactions of treatment receipt in equation (8) and the results support their inclusion in four of the six specifications in grades two and three.

[34]We only use one lagged test score to identify the attrition equation. Thus, we do not require individuals to have completed exams in all three cognitive subject areas.

[35]We consider the nonparametric series estimator proposed in Hirano et al. (2003). In implementation we considered using up to a third order and then used the AIC criterion to determine which terms should remain in the specification.

References

[1] Angrist, Joshua D.; Imbens, Guido W. and Rubin, Donald B. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association,* 1996, 91(434), pp. 444-55.

[2] Balke, Alexander; and Pearl, Judea. "Bounds on Treatment Effects from Studies With Imperfect Compliance." *Journal of the American Statistical Association,* 1997, 92, pp. 1171-1177.

[3] Barnard, John; Du, Jiangtao; Hill, Jennifer L. and Rubin, Donald B. "A Broader Template for Analyzing Broken Randomized Experiments." *Sociological Methods & Research*, 1988, 27(2), pp. 285-317.

[4] Becketti, Sean; Gould, William; Lillard, Lee and Welch Finis. "The Panel Study of Income Dynamics after Fourteen Years: An Evaluation." *Journal of Labor Economics,* 1988, 6(4), pp. 472-92.

[5] Ben-Porath, Yoram. "The Production of Human Capital and the Life-Cycle of Earnings." *Journal of Political Economy*, August, 1967.

[6] Boardman, Anthony E. and Murnane, Richard J. "Using Panel Data to Improve Estimates of the Determinants of Educational Achievement." *Sociology of Education,* 1979, 52, pp. 113-121.

[7] Chesher, Andrew "Identification in Nonseparable Models," *Econometrica*, 2003, 71, pp. 1405-1441.

[8] Coleman, James S. "Some Points on Choice in Education." *Sociology of Education*, 1992, 65(4), pp. 260-2.

[9] Ding, Weili and Lehrer, Steven F. "Estimating Dynamic Treatment Effects and Bounds for the Average Treatment Effect from Project STAR," Working Paper, Queen's University, 2008.

[10] Ding, Weili and Lehrer, Steven F. "Accounting for Time-Varying Unobserved Ability Heterogeneity within Education Production Functions." Working Paper, Queen's University, 2004.

[11] DuMouchel, William H. and Duncan, Greg J. "Using Sample Survey Weights in Multiple Regression Analyses of Stratified Samples." *Journal of the American Statistical Association,* 1983, 78(383), pp. 535-43.

[12] Finn, Jeremy D.; Gerber, Susan B.; Achilles, Charles M. and Boyd-Zaharias, Jayne. "The Enduring Effects of Small Classes." *Teachers College Record*, 2001, 103(2), pp. 145-83.

[13] Fitzgerald, John; Gottschalk, Peter and Moffitt, Robert. "An Analysis of Sample Attrition in Panel Data: The Michigan Panel Study of Income Dynamics." *Journal of Human Resources,* 1998, 33(2), pp. 300-44.

[14] Frangakis, Costas E. and Rubin, Donald B. "Principal stratification in causal inference." *Biometrics*, 2002, 58(1), pp. 21-9.

[15] Frangakis, Costas E. and Rubin, Donald B. "Addressing complications of intention-to-treat analysis in the presence of all-or-none treatment-noncompliance and subsequent missing outcomes." *Biometrika,* 1999, 86(2), pp. 365-79.

[16] Hanushek, Eric A. "Some Findings from an Independent Investigation of the Tennessee STAR Experiment and from Other Investigations of Class Size Effects." *Educational Evaluation and Policy Analysis*, 1999, 21(2), pp. 143-63.

[17] Heckman, James J.; Lalonde, Robert and Smith, Jeffrey. "The Economics and Econometrics of Active Labor Market Programs," in Orley Ashenfelter and David Card, eds., *Handbook of Labor Economics, Volume 3*, Amsterdam: Elsevier Science, 2001.

[18] Heckman, James J.; Hohmann, Neil, Khoo, Michael and Smith Jeffrey. "Substitution and Dropout Bias in Social Experiments: A Study of an Influential Social Experiment." *Quarterly Journal of Economics*, 2000, 115(2), pp. 651-90.

[19] Heckman, James J.; Smith, Jeffrey and Taber, Chris. "Accounting For Dropouts in the Evaluation of Social Experiments." *Review of Economics and Statistics,* 1998, 80(1), pp. 1-14.

[20] Hirano, Keisuke; Imbens, Guido and Ridder, Geert. "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score." *Econometrica*, 2003, 71(5), pp. 1161-89.

[21] Horowitz, Joel L.; and Manski, Charles F. "Identification and Robustness with Contaminated and Corrupted Data." *Econometrica,* 1995, 63(2), pp. 281-302.

[22] Hoxby, Caroline M. "The Effects of Class Size on Student Achievement: New Evidence from Population Variation." *Quarterly Journal of Economics*, 2000a 115(4), pp. 1239-85.

[23] Hoxby, Caroline M. "Peer Effects in the Classroom: Learning from Gender and Race Variation" Peer Effects in the Classroom: Learning from Gender and Race Variation." *NBER Working Paper No. W7867,* 2000b.

[24] Imbens Guido W. and Angrist, Joshua D. "Identification and Estimation of Local Average Treatment Effects." *Econometrica,* 1994, 62(2) pp. 467–76.

[25] Krueger, Alan B. "Experimental Estimates of Education Production Functions." *Quarterly Journal of Economics,* 1999, 114(2), pp. 497-532.

[26] Lechner, Michael. "Sequential Matching Estimation of Dynamic Causal Models." Working Paper, University of St. Gallen, 2004.

[27] Lechner, Michael and Miquel, Ruth. "Identification of Effects of Dynamic Treatments by Sequential Conditional Independence Assumptions." Working Paper, University of St. Gallen, 2005.

[28] Manski, Charles F. "Nonparametric Bounds on Treatment Effects." *American Economic Review*, 1990, 80(2), pp. 319-23.

[29] Miquel, Ruth. "Identification of Effects of Dynamic Treatments with a Difference-in-Differences Approach," Working Paper, University of St. Gallen, 2003.

[30] Miquel, Ruth. "Identification of Dynamic Treatment Effects by Instrumental Variables." Working Paper, University of St. Gallen, 2002.

[31] Newey, Whitney K. "A Method of Moments Interpretation of Sequential Estimators." *Economics Letters,* 1984, 14, pp. 201-206.

[32] Robins James M.; Miguel A. Hernán and Babette Brumback. Marginal Structural Models and Causal Inference in Epidemiology. *Epidemiology,* 2000, 11, pp. 550–560

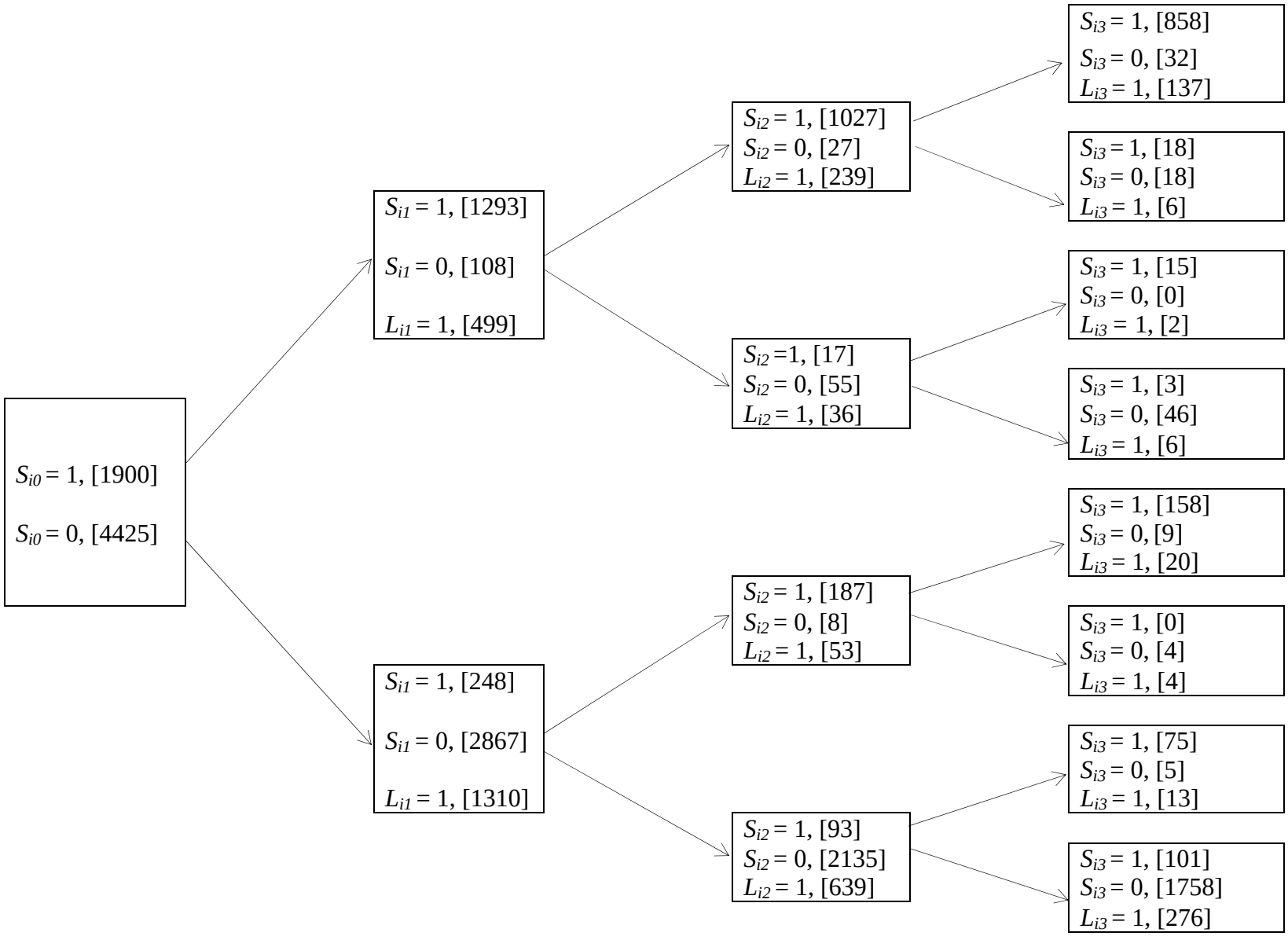
[33] Robins, James M. "A New Approach to Causal Inference in Mortality Studies with Sustained Exposure Periods - Application to Control of the Healthy Worker Survivor

Effect," *Mathematical Modelling*, 1986, 7, pp. 1393-1512, with 1987 Errata, *Computers and Mathematics with Applications,* 14, pp. 917-21; 1987 Addendum, *Computers and Mathematics with Applications*, 14, pp. 923-45; and 1987 Errata to Addendum, *Computers and Mathematics with Applications*, 18, pp. 477.

[34] Wooldridge, Jeffrey M. "Inverse Probability Weighted M-estimators for Sample Selection, Attrition and Stratification." *Portuguese Economic Journal,* 2002, 1(2), pp. 117-39.

[35] Yau, Linda and Little, Roderick J. A. "Inference for the Complier-Average Causal Effect from Longitudinal Data Subject to Noncompliance and Missing Data, with Application to a Job Training Assessment for the Unemployed." *Journal of the American Statistical Association,* 2001, 96(456), pp. 1232-44.

Figure 1: Transitions During Project Star for Kindergarten Cohort



Note: Number of individuals are in [ ] parentheses.

Table 1: Testing Randomization of Student Characteristics across Class Types

	Kindergarten	Grade One	Grade Two	Grade Three
INCOMING STUDENTS				
White or Asian Student	2.35*10E-4 (0.012)	-0.275* (0.193)	-0.061* (0.041)	7.63*10E-4 (0.063)
Female Student	0.012 (0.019)	0.199* (0.126)	-0.020 (0.021)	-0.017 (0.028)
Student on Free lunch	-8.74*10E-3 (0.017)	-0.262* (0.167)	0.013 (0.022)	-0.057* (0.037)
Joint Test of Student Characteristics	0.29 [0.831]	1.83* [0.150]	1.24 [0.301]	1.01 [0.392]
Number of Observations	6300	2211	1511	1181
R Squared	0.318	0.360	0.248	0.411
FULL SAMPLE				
White or Asian Student	2.35*10E-4 (0.012)	-0.003 (0.021)	-0.008 (0.025)	-0.021 (0.027)
Female Student	0.012 (0.019)	0.007 (0.009)	0.004 (0.009)	0.008 (0.009)
Student on Free lunch	-8.74*10E-3 (0.017)	-0.038*** (0.016)	-0.030** (0.016)	-0.044*** (0.016)
Joint Test of Student Characteristics	0.29 [0.831]	2.05* [0.114]	1.38 [0.255]	2.98*** [0.037]
Number of Observations	6300	6623	6415	6500
R Squared	0.318	0.305	0.328	0.359

Note:Regressions include school indicators. Standard errors corrected at the school level are in ( ) parentheses. Probability > F are in [ ] parentheses. ***,**,* indicate statistical significance at the 5%, 10% and 20% level respectively.

.

Table 2: Are Attritors Different from Non-attritors?

Subject Area	Mathematics	Reading	Word Recognition
Kindergarten Class Type	10.434***	6.513***	7.370***
	(2.332)	(1.440)	(1.628)
White or Asian Student	20.499***	8.608***	8.505***
	(2.760)	(2.005)	(2.524)
Female Student	2.587**	3.349***	2.488**
	(1.363)	(1.074)	(1.296)
Student on Free lunch	-13.729***	-12.239***	-13.916***
	(1.679)	(1.187)	(1.480)
Years of Teaching Experience	0.323*	0.255***	0.329***
	(0.220)	(0.123)	(0.135)
White Teacher	-0.926	-1.577	-1.578
	(4.366)	(3.068)	(3.506)
Teacher has Master Degree	-1.482	-1.211	-0.491
	(2.396)	(1.423)	(1.729)
Attrition Indicator	-17.305***	-13.674***	-13.198***
	(3.838)	(2.537)	(3.251)
Attrition Indicator Interacted with	-5.383***	-2.069	-3.004*
Kindergarten Class Type	(2.616)	(1.686)	(2.045)
Attrition Indicator Interacted with	-3.949*	-.259	-1.177
White or Asian Student	(2.732)	(1.824)	(2.368)
Attrition Indicator Interacted with	5.597***	2.943***	3.750***
Female Student	(2.078)	(1.454)	(1.739)
Attrition Indicator Interacted with	-5.186***	-0.496	0.549
Student on Free lunch	(2.384)	(1.554)	(1.891)
Attrition Indicator Interacted with	0.188	0.075	-0.060
Years of Teaching Experience	(0.210)	(0.131)	(0.164)
Attrition Indicator Interacted with	1.263	2.269	0.642
White Teacher	(3.490)	(2.133)	(2.678)
Attrition Indicator Interacted with	-1.370	0.939	1.552
Teacher has Master Degree	(2.490)	(1.586)	(1.876)
Number of Observations (R-Squared)	5810 (0.305)	5729 (0.295)	5789 (0.259)
Joint Effect of Attrition on Constant	42.39***	32.68***	25.76***
and Coefficient Estimates	[0.000]	[0.000]	[0.000]
Joint Effect of Attrition on all	3.14***	1.23	1.45*
Coefficient Estimates but not constant	[0.003]	[0.280]	[0.181]
Effect of Attrition	20.33***	29.06***	16.48***
on Constant Alone	[0.000]	[0.000]	[0.000]

Note: Regressions include school indicators. Standard errors corrected at the classroom level are in ( ) parentheses. Probability > F are in [ ] parentheses. ***,**,* indicate statistical significance at the 5%, 10% and 20% level respectively.

Table 3: Structural Estimates of the Treatment Parameters in Education Production Functions

Subject Area	Mathematics	Reading	Word Recognition
Kindergarten			
S_{iK}	8.595 (1.120)***	5.950 (0.802)***	6.342 (0.945)***
Grade One			
S_{iK}	7.909 (4.625)**	8.785 (5.284)**	11.868 (6.722)**
S_{i1}	9.512 (3.307)***	9.315 (4.350)***	15.394 (5.730)***
$S_{iK}S_{i1}$	-6.592 (5.648)	-2.229 (6.992)	-11.060 (8.965)
Grade Two			
S_{iK}	-2.078 (7.276)	11.320 (7.240)*	9.959 (8.438)
S_{i1}	-4.010 (3.855)	-20.036 (19.189)	4.298 (7.763)
S_{i2}	15.150 (5.430)***	3.040 (4.428)	0.526 (5.814)
$S_{iK}S_{i1}$	3.851 (11.678)	1.148 (24.059)	-12.074 (17.673)
$S_{iK}S_{i2}$	-4.049 (13.112)	-31.513 (17.366)**	-23.084 (13.237)**
$S_{i1}S_{i2}$	-4.944 (6.617)	25.122 (19.480)*	7.868 (8.537)
$S_{iK}S_{i1}S_{i2}$	6.653 (16.067)	23.634 (28.632)	30.111 (19.851)*
Grade Three			
S_{iK}	-7.298 (10.901)	1.215 (10.372)	13.071 (12.202)
S_{i1}	43.514 (32.898)*	22.083 (30.097)	-6.920 (37.200)
S_{i2}	25.263 (42.080)	-22.085 (26.069)	-25.024 (22.031)
S_{i3}	-6.835 (3.932)**	-10.590 (4.179)***	-12.738 (5.952)***
$S_{iK}S_{i1}$	-38.612 (30.944)	7.978 (39.071)	-18.002 (32.872)
$S_{iK}S_{i2}$	37.355 (28.625)*	-42.740 (25.731)**	-2.932 (22.527)
$S_{iK}S_{i3}$	-39.819 (19.922)***	17.870 (18.147)	7.328 (14.855)
$S_{i1}S_{i2}$	-61.947 (52.749)	25.388 (35.964)	-7.586 (36.814)
$S_{i1}S_{i3}$	17.163 (43.057)	-6.613 (32.183)	-7.954 (29.718)
$S_{i2}S_{i3}$	-14.366 (42.280)	35.547 (22.836)*	29.203 (26.267)
$S_{iK}S_{i1}S_{i3}$	-4.651 (52.881)	-41.180 (43.335)	-14.706 (35.985)
$S_{iK}S_{i1}S_{i2}S_{i3}$	48.084 (48.704)	6.834 (30.521)	14.377 (33.920)

Note: Corrected standard errors in parentheses. The sequences $S_{iK}S_{i1}S_{i2}$, $S_{iK}S_{i2}S_{i3}$ and $S_{i1}S_{i2}S_{i3}$ lack unique support to permit identification in grade 3. ***,**,* indicate statistical significance at the 5%, 10% and 20% level respectively.

Table 4: Dynamic Average Treatment Effect for the Treated Estimates

Subject Area	Mathematics	Reading	Word Recognition
Kindergarten			
$\tau^{(1)(0)}(1)$	8.595 (1.120)***	5.950 (0.802)***	6.342 (0.945)***
Grade One			
$\tau^{(0,1)(0,0)}(0,1)$	9.512 (3.307)***	9.315 (4.350)***	15.394 (5.730)***
$\tau^{(1,0)(0,0)}(1,0)$	7.909 (4.625)**	8.785 (5.284)**	11.868 (6.722)**
$\tau^{(1,1)(0,0)}(1,1)$	10.829 (8.021)*	15.872 (9.787)*	16.203 (12.587)*
$\tau^{(1,1)(1,0)}(1,1)$	2.920 (6.544)	7.086 (8.235)	4.334 (10.640)
$\tau^{(1,1)(0,1)}(1,1)$	1.317 (7.300)	6.556 (8.764)	0.808 (11.205)
$\tau^{(0,1)(1,0)}(0,1)$	1.603 (5.686)	0.530 (6.844)	4.066 (8.833)
Grade Two			
$\tau^{(0,0,1)(0,0,0)}(0,0,1)$	15.150 (5.430)***	3.040 (4.428)	0.526 (5.814)
$\tau^{(1,0,0)(0,0,0)}(1,0,0)$	-2.078 (7.276)	11.320 (7.240)*	9.959 (8.438)
$\tau^{(1,1,1)(0,0,0)}(1,1,1)$	10.574 (26.606)	12.714 (50.199)	17.603 (33.463)
$\tau^{(1,1,1)(1,0,0)}(1,1,1)$	12.651 (25.589)	1.394 (49.674)	7.644 (32.381)
$\tau^{(1,1,1)(1,1,0)}(1,1,1)$	12.810 (22.436)	20.282 (38.993)	15.421 (25.999)
$\tau^{(0,1,1)(0,0,0)}(0,1,1)$	6.196 (9.400)	8.125 (27.700)	12.691 (12.920)
$\tau^{(0,0,1)(1,0,0)}(0,0,1)$	17.228 (9.084)**	-8.208 (8.490)	-9.433 (10.249)
Grade Three			
$\tau^{(0,0,0,1)(0,0,0,0)}(0,0,0,1)$	-6.835 (3.932)**	-10.590 (4.179)***	-12.738 (5.952)***
$\tau^{(1,1,1,1)(0,0,0,0)}(1,1,1,1)$	-2.148 (129.436)	-17.192 (93.135)	-20.985 (102.228)
$\tau^{(1,1,1,1)(1,1,0,0)}(1,1,1,1)$	0.247 (120.810)	-22.487 (81.117)	-35.114 (85.973)
$\tau^{(1,1,1,1)(1,1,1,0)}(1,1,1,1)$	-0.424 (96.033)	10.115 (63.543)	7.262 (70.360)
$\tau^{(1,1,1,1)(0,1,1,1)}(1,1,1,1)$	-4.940 (86.378)	-20.263 (64.365)	-30.626 (75.468)
$\tau^{(0,1,1,1)(0,0,0,0)}(0,1,1,1)$	2.792 (96.397)	3.071 (67.314)	9.641 (68.958)
$\tau^{(0,0,1,1)(0,0,0,0)}(0,0,1,1)$	4.062 (59.781)	-3.472 (37.243)	-2.215 (32.284)
$\tau^{(0,0,1,1)(1,1,0,0)}(0,0,1,1)$	6.458 (75.714)	-8.767 (59.001)	-16.344 (64.043)
$\tau^{(1,1,0,0)(0,0,0,0)}(1,1,0,0)$	2.396 (46.461)	-7.568 (31.614)	2.396 (46.461)

Note: Standard Errors in parentheses.
***,**,* indicate statistical significance at the 5%, 10% and 20% level respectively.

Table 5: Tests of Weighted versus Unweighted Estimates

Subject Area	Mathematics	Reading	Word Recognition
Grade One	8.74 [0.000]	3.39 [0.000]	1.35 [0.169]
Grade Two	1.48 [0.071]	3.86 [0.000]	2.08 [0.002]
Grade Three	1.72 [0.008]	1.91 [0.002]	1.03 [0.424]

Note: Probability > F are in [ ] parentheses.

Table 6: Likelihood Ratio Tests for the Presence of Selection on Unobservables

Subject Area	Mathematics	Reading	Word Recognition
Grade One	2661.91 [0.000]	4468.98 [0.000]	3293.98 [0.000]
Grade Two	1648.11 [0.000]	1478.86 [0.000]	5480.28 [0.000]
Grade Three	1606.95 [0.000]	1421.94 [0.000]	839.84 [0.000]

Note: Probability > χ^2 are in [ ] parentheses